Session 3:

Experimental design = Erors Avoiding errors

Understanding errors & Experimental design

- Understanding different sources of errors
  - Systematic errors
  - Statistical errors
    - Type I errors
    - Type II errors
- Experimental design: avoiding, correcting, quantifying errors
  - Independent evidence from alternative methods
  - Replication, Replication, Replication
  - Randomization, iodmzoanaintR, antRoomdnziia
  - Controls and Initial conditions
  - Planning experiments

Noise

- Ion channel open (high current) or closed (low current)

- Snow on an old TV screen

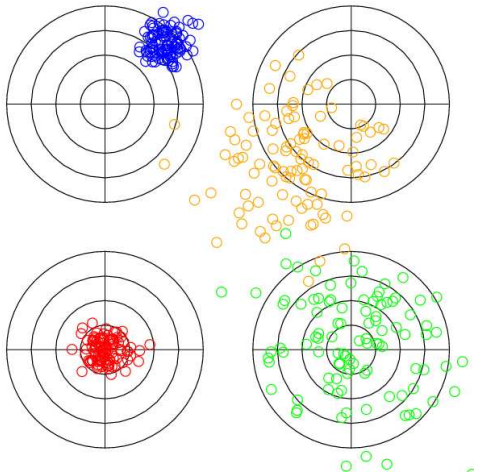
The two main types of errors

- Systematic errors
  - Bias in sampling
  - Bias in measurement method
  - Malfunctioning equipment
  - Wrong equipment settings
- Systematic errors reduce accuracy
- How to deal with:
  - Randomize sampling to avoid bias
  - Use controls (with known properties) to detect
- Statistical errors
  - Noise in measurements
  - Random fluctuations
  - Variation in sample
- Reduce precision of your method
- Can be quantified if making replicates
- How to deal with:
  - Make more measurements
  - Decrease variation by choosing more uniform sample

Examples of errors

- Systematic errors
  - Bias in sampling: using a phonebook for polling in 1920
  - Bias in method: sugar assay detects some sugars with higher sensitivity than others
  - Photometer set to wrong wavelength
  - pH meter calibrated with spoilt buffers
- Statistical errors
  - Noise, fluctuation, random variation
- Make sample more uniform (reduce variation)
  - Single age group
  - Blocking a field site
  - Grow cells under well controlled and reproducible conditions

Precise? Accurate?



Statistical errors

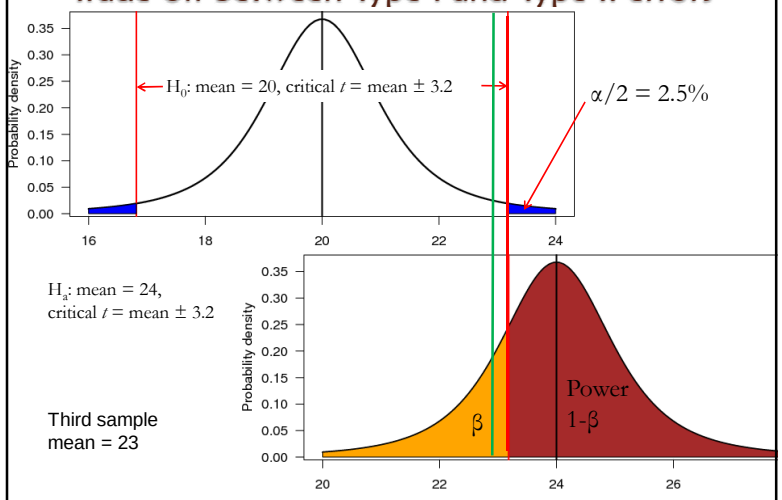
Two types of statistical errors

- Type I error = α error = false positive rate
  - Reject null hypothesis when it is true
- Type II error = β error = false negative rate
  - Accept null hypothesis when it is false
- Statistical tests involve a trade-off between α error and β error
- Compromise often used (though arbitrary): α=0.05, β=0.2

Trade-off between Type I and Type II errors

- Consider this scenario
  - We don't know the true (population) mean but have taken two small samples (n = 4) from this population with these results
    - Sample A: mean = 20
    - Sample B: mean = 24
  - Therefore we hypothesize that the population mean is
    - H_0 (null-hypothesis): true mean = 20
    - H_a (alternative hypothesis): true mean = 24
  - Based on a t-test (used to test whether sample means are the same), we can work out the critical *t* value for rejecting these hypotheses (with n = 4, α = 5 %) to be *t* = 3.2

Trade-off between Type I and Type II errors



Trade-off between Type I and Type II errors

- From this rather constructed but possible example we see
  - The smaller we make the α error, the larger the β error gets, the smaller the power of the test
  - Trade-off between α and β
  - Power would increase with
    - larger distance between the two means (larger effect size)
    - smaller variance of the data (sharper peaks)
    - larger sample size (we are less likely to deviate from the true mean)
    - larger α
    - choosing a test with higher power that can be used if the data fulfil stricter assumptions

Specificity versus Sensitivity

Statistical errors

		Known condition: Positive (Present)	Known condition: Negative (Absent)
Test outcome	Positive	True Positive (TP)	False Positive (FP) (Type I error)
	Negative	False Negative (FN) (Type II error)	True Negative (TN)

false negative rate $\beta = \frac{FN}{\text{all P}} = \frac{FN}{TP + FN}$

false positive rate $\alpha = \frac{FP}{\text{all N}} = \frac{FP}{TN + FP}$

sensitivity $= \frac{TP}{\text{all P}} = \frac{TP}{TP + FN} = \frac{\text{all P - FN}}{\text{all P}} = 1 - \beta$

specificity $= \frac{TN}{\text{all N}} = \frac{TN}{TN + FP} = \frac{\text{all N - FP}}{\text{all N}} = 1 - \alpha$

sensitivity = power

Statistical errors

- Example: HIV screening
  - "A large study of HIV testing in 752 U.S. laboratories reported a sensitivity of 99.7% and specificity of 98.5% for enzyme immunoassay" (Chou *et al.* (2005). Screening for HIV: a review of the evidence for the U.S. preventive services task force. *Annals of Internal Medicine* (143): 55-73)
  - What we know:
    - specificity = 0.985 ⇒ false positive rate $\alpha = 1 - 0.985 = 0.015$
    - sensitivity = 0.997 ⇒ false negative rate $\beta = 1 - 0.997 = 0.003$
      - At a false positive rate of 1.5% and a false negative rate of 0.3% this is not a bad test for screening purposes
    - UK population = 61,186,000 (total)
    - HIV positive = 73,000 (condition present: known to be HIV positive)

Statistical errors

		Known condition: HIV^+ = all P = 73,000	Known condition: HIV^- = all N = total – all P = 61,113,000
Test outcome	Positive	TP = all P – FN = 72,781	FP = α * all N = 916,695
	Negative	FN = β * all P = 219	TN = all N – FP = 60,196,305

- HIV screening of the UK population will give loads of false positives although the specificity of the test is good
- General problem in screening for some rare condition
  - Extreme case: whole population negative (condition absent)
- Opposite if screening for something common
  - Extreme case: whole population positive (condition present)

Exercise

		Known condition (as confirmed by endoscopy): Positive	Known condition (as confirmed by endoscopy): Negative
Test outcome	Positive	TP = 2	FP = 18
	Negative	FN = 1	TN = 182

- The Fecal Occult Blood (FOB) test is used in screening for bowel cancer
- From the above results of a clinical study, calculate the
  - false positive rate
  - specificity
  - false negative rate
  - sensitivity
  - power
- Is this test satisfactory for screening?

Independent Evidence

The types of errors we make

- **Check for systematic errors and make enough replicates**
  - Nevertheless, it is always possible that you fail to detect a problem with your instrumentation, that you reject the null hypothesis when it is true, or that you accept the null hypothesis when it is false
  - The best way of making sure your conclusions are correct is by supporting them with independent evidence using alternative methods
  - Getting the same results via entirely independent routes is more valuable than getting the same results from two independent measurements using the same protocol
  - Stats alone is not enough

Independent evidence

- Example 1
  - Measure pH with glass electrode AND pH indicators: do both measures agree?
- Exercise: think of further examples...

Independent evidence

- The gold standard: a theory that makes many predictions that can be independently tested
  - A reliable theory is supported by lots of independent evidence: it can explain a large variety of different phenomena
  - Newton's laws of mechanics predict
    - the swinging of a pendulum
    - the acceleration of a falling apple
    - the orbits of the planets
    - the tides
    - how you can turn the front wheel when cycling no hands by leaning to a side
    - even postulating the existence of planet Neptune to explain the discrepancies between predicted and observed orbits of Uranus

Replication

Replication

- Why Replication?
  - Increases reliability
  - Quantifies variability
- How much Replication?
  - Rule of thumb: 30 replicates
- Can we calculate how many we really have to do?

Replication

- We can calculate this!
  - Assuming you are going to use a *t*-test to compare the mean between two samples or treatments, and that the errors are normally distributed etc.
  - Let δ be the difference between two sample means you want to be able to detect (effect size)

δ = *t*-statistic · standard error of the difference of the means (s.e.d.m.)

(eq. 1) $$\delta = t\sqrt{\frac{2s^2}{n}}$$

  - Solve for number of replicates *n*

(eq. 2)

Replication

- Example: From a pilot study, we estimate mean = 50 and SD = 20
- Suppose we want to be able to detect a small effect as significant, say a difference of 10% from this mean, i.e. δ = 5
- Calculate *n* using the R function power.t.test()

```
> power.t.test(delta=5,sd=20,sig.level=0.05,
power=0.8,type='one.sample')

     One-sample t test power calculation

              n = 127.5161
          delta = 5
             sd = 20
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Replication

- Number of replicates *n*, power (1-β), significance level α, standard deviation *s*, and effect size δ are all coupled (each variable is a function of the others)
- You can calculate one, if all others are known or given
- Calculate *n*

```
> power.t.test(delta=5,sd=20,sig.level=0.05,
  power=0.8,type='one.sample')
```

- Calculate δ

```
> power.t.test(n=128,sd=20,sig.level=0.05,
  power=0.8,type='one.sample')
```

- Calculate power (that's where the name comes from), a.k.a. sensitivity

```
> power.t.test(delta=5,n=128,sd=20,sig.level=0.05,
  type='one.sample')
```


Exercise: replication

- From previous studies you know that a control group of untreated plants has a yield of 100 ± 10 kg m^{-2} (mean ± standard deviation)
  - Using the standard significance level of 0.05 and power of 0.80, calculate the sample size required to detect a difference of 10% from the mean of the control group as significant

What is a sample?

- Note different meanings of 'sample'
  - Statisticians use the word sample for a subset of the whole population
    - e.g. 100 individual diabetes patients are a sample of the UK diabetes population
  - Experimentalists use sample for a single item
    - e.g. a blood sample from one individual patient at one particular time

What is a replicate?

- Pseudoreplicates
  - Replicates that aren't true, independent replicates
  - Pretend a high *n*
    - Measuring glucose concentration in different aliquots (portions) of the same blood 'sample'
      - this is a technical, but not biological replicate
    - Measuring glucose concentration in different samples from the same patient collected every day at the same time
      - appropriate only if the variation for a single individual is what you want to know
    - Measuring glucose concentration in different individuals
      - this is what you would usually call a biological replicate, encompassing all the biological variation and including technical variation

Pseudoreplication

- You have to expect temporal or spatial autocorrelation in such data
- That means your data points are not independent and independence of errors is a crucial assumption behind the standard statistical tests
- What you can do (using temporally correlated data as an example)
  - Average over all time points and perform analysis on the means: you loose information by averaging
  - Perform separate analysis for each time point: ignores dependencies
  - Filter out autocorrelation or correct for autocorrelation
  - Analyse autocorrelated data with proper time series analysis
  - For spatial autocorrelation there is geostatistics

Pseudoreplication retake

- Actually, if it's pseudoreplication or not depends on the question you want to answer, so let's revisit our first example
  - Measuring glucose concentration in several portions of the same blood 'sample'
    - This would be OK (true replicates) *iff* you want to know how precise your glucose measurements are rather than the variation of glucose levels among patients
  - Measuring glucose concentration in different samples from the same patient collected every day at the same time
    - This would be OK *iff* you want to know the variation of an individual's glucose concentration over time
  - You might actually want to know the variation on all three levels
    - Within sample ⇒ measurement error
    - Individual patient over time ⇒ temporal variability
    - Between patients ⇒ variability in population

Exercise

- To estimate number of bacteria in a sample, you make a dilution series and plate out 3 replicate aliquots of each dilution
  - Correct procedure?

Randomization

Randomization

- We randomize to reduce bias (systematic error)
- Proper randomization is not always easy
- Example: picking trees in a forest at random
  - Idea: generate random locations (map coordinates) and then pick the trees closest to the randomly chosen positions (which are unlikely to have hit upon a tree)
  - Let's try it out

Randomization

- So that's not proper randomization
- The only way to do it is by actually numbering all the trees in the forest (enjoy) and then randomly picking from the list of tree numbers
- Make a random permutation of the sequence of numbers
  - Like shuffling cards
  - In R you can use sample() for shuffling

```
> sample(1:number_of_trees)
```

- sample() can do more than shuffling, you can randomly pick n elements from vector x with/without replacement like this:
  - sample(vector, num_elements, replace = FALSE)

```
> sample(1:10, 4, replace=FALSE)
[1] 9 6 4 8
```


Randomization

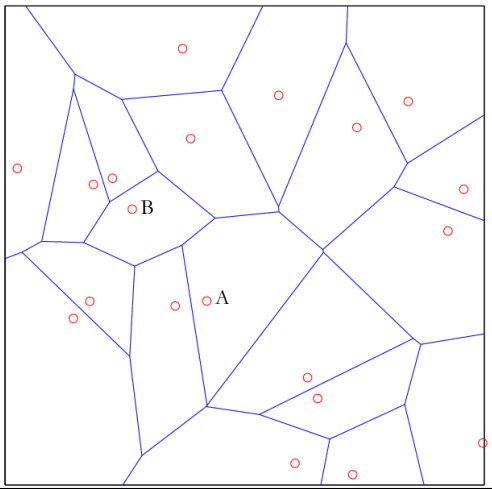
The neighbourhood of tree A

(the set of locations closer to A than any other tree, also known as the Voronoi neighbourhood)

is larger than that of tree B

so tree A would be selected with higher probability!

Voronoi cells of 20 randomly located trees



Exercise: randomization

- The task is to randomly assign 100 patients to 2 groups that will receive different treatments (drug or placebo)
  - The patient id's are numbered from 1 to 100 for convenience
  - Make a vector **p** containing all 100 patient id's
  - Make a random permutation of these 100 numbers (that means all patient id's should occur exactly once) and store them in a new variable **pr**
  - Assign the first half of **pr** to treatment A and the second to treatment B

Controls

- Negative controls
  - Leave out test substance
- Positive controls
  - Include known amounts of known substance = standard
  - Internal standard
  - External standard
  - Good to detect systematic errors
- Calibration curve
  - Covers a range of concentrations from 0 to x
  - Includes negative and positive controls

Initial conditions

- Initial conditions in different samples (in the experimental sense) are likely different (whether you have picked the samples randomly, which you should, or not)
- Therefore you have to measure them
  - Take aliquots at time 0
  - Always take a measurement before starting treatment including control treatment

Some practical advice on experiments

- Measuring change over time (e.g. rates of reactions, kinetics)
  - If you don't know how much activity you can expect over a range of treatments, or if you have both very fast and very slow reactions in your assay...
    - ...then use a log scale for your time points such as this
    - 0, 1, 2, 4, 8, 16, 32, 64, 128, ... min
    - of course you must include time zero in your sampling, as initial conditions can be different from tube to tube

Some practical advice on experiments

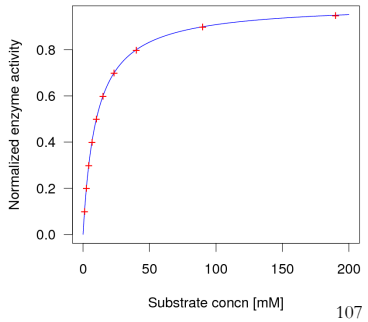
- Enzyme kinetics
  - If you have a rough idea of K_m and V_{max} from pilot data...
  - ...then you can calculate the substrate concentration to use to get a good coverage of activities v/V_{max} (.1 .2 .3 .4 .5 .6 .7 .8 .9 .95)

Michaelis-Menten kinetics

$$v = \frac{V_{max}\, s}{K_m + s}$$

solved for substrate concentration s

$$s = \frac{v/V_{max}\, K_m}{1 - v/V_{max}}$$



Some practical advice on experiments

- More generally, if your response is not linear, making measurements at equidistant points on the x axis is simple but not smart

Measurement scales

Quantitative or continuous variables	Ratio	Zero value defined so can calculate proportions	Concentrations, energy, lengths, absolute temperature Makes sense to say 'twice as much'
	Interval	Distances defined so can calculate differences	Temperature in Celsius, calendar dates
Categorical variables	Ordinal	Rank order defined	Low, medium, high (you can score, rate or rank)
	Nominal	Distinctions defined	Bird, Mammal; Female, Male (you can classify but not rank)

Don't calculate sums, differences, means, etc.!