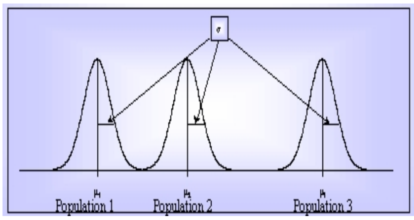
Statistics and Probability

Power; 2-sample tests; ANOVA



http://www.isrec.isb-sib.ch/~darlene/geneve/

Brief review of hypothesis testing

- 2 'competing theories' regarding a population parameter:
  - *NULL* hypothesis *H* ('straw man')
  - *ALTERNATIVE* hypothesis *A* ('claim', or theory you wish to test)
- *H:* NO DIFFERENCE
  - any observed deviation from what we expect to see is due to *chance variability*
- *A:* THE DIFFERENCE IS *REAL*

Test statistic

- Measure how far the observed data are from what is expected *assuming the NULL H* by computing the value of a *test statistic* (TS) from the data
- The particular TS computed depends on the parameter
- For example, to test the population mean μ, the TS is the sample mean (or standardized sample mean)
- In R: `> t.test(x)`

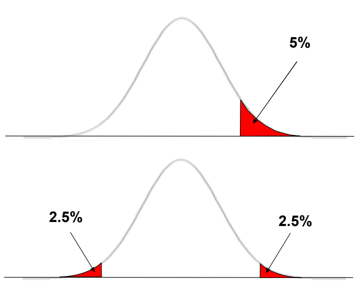
One-sided vs. two-sided test

- The choice of alternative influences the conclusion
- If the *ALT* is "the coin is biased", we do not specify the direction of the bias
- This is a *two-sided test*
- If α is *e.g.* 0.05, it means we must allow $\alpha/2$ (0.025) for bias towards tail and $\alpha/2$ (0.025) for bias towards head
- If the *ALT* is "the coin is biased towards heads", we specify the direction of the bias and the test is *one-sided*

One-sided vs two-sided test

One-sided
e.g. A: μ>0

Two-sided
e.g. A: μ=0



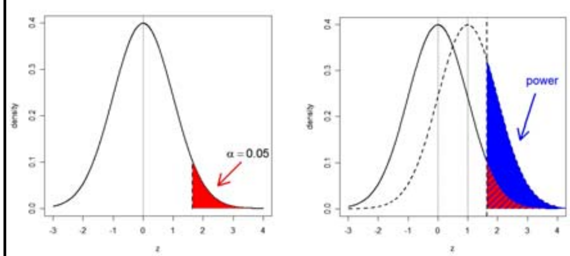
Example

- Are UniGE students healthier than average? Assume that on average, students in Switzerland exercise 3 hours per week. A survey of 64 UniGE students yields a sample mean of 3.4 hours with standard deviation 1.6 hours.
- At α = 1%, is there statistical evidence that UniGE students exercise more*??*

Power

- Not only do you want to have a low FALSE positive rate, but you would also like to have a high TRUE positive rate – that is, high *power*, the chance to find an effect (or difference) if it is really there
- Statistical tests will not be able to detect a true difference if the *sample size* is too small compared to the *effect size* of interest
- In R: `> power.t.test(x)`

Power example



Power and sample size

- In applications, power calculations under *different scenarios* (varying effect size, SD, α) are often used to decide on sample size when *planning experiments*
- Choose effect size(s) of interest to detect
- Would like to have high probability of detecting a true difference of this size (or larger) -> want high power
- Often use power of .75, .8, .9
- *Higher power* can be achieved for fixed α (and effect size) by *increasing sample size*

Back to example

- Are UniGE students healthier than average?
- We carried out a test of H: $\mu = 3$
- Suppose that, unknown to us, the true UniGE mean is actually 3.4.
- What is the power of the test at this particular alternative*??*

R: power example

```
> power.t.test(n=64, delta=.4, sd=1.6,
type= "one.sample",alternative="one.sided")

     One-sample t test power calculation

              n = 64
          delta = 0.4
             sd = 1.6
      sig.level = 0.05
          power = 0.6306582
    alternative = one.sided
```


R: sample size example

```
> power.t.test(power=.80, delta=.4, sd=1.6,
type= "one.sample",alternative="one.sided")

     One-sample t test power calculation

              n = 100.2877
          delta = 0.4
             sd = 1.6
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

Trees

- A study is conducted to investigate the growth of a certain type of tree at an elevation of 675 meters
- The variable of interest is the core measurement (in cm) for a 10 year period
- The theory is that the mean should be at least 1.75
- In a random sample of 10 measurements, the mean was 2 with an SD of .5

Steps in hypothesis testing (I)

1. Identify the population parameter being tested

   Here, the parameter being tested is the population mean core measurement μ

2. Formulate the NULL and ALT hypotheses

   H: μ = 1.75 (or $\mu \leq 1.75$)

   A: $\mu > 1.75$

3. Compute the TS

   $t = (2- 1.75)/(.5/\sqrt{10}) = 1.58$

Steps in hypothesis testing (II)

4. Compute the p-value

   Here, $p = P(T_9 > 1.58) = .07$

5. (Optional) *Decision Rule:* REJECT H if the p-value $\leq \alpha$

   If we use α = .05, the decision here will be DO NOT REJECT H (but just barely!)

More trees

- Now say we are interested in whether the mean core measurement is the same in trees at 675 meters and trees at 825 meters
- Now say we have a random sample also of size 10 of trees at 825 meters, with a mean core measurement of 2.65 cm and SD 1.15 cm
- How might we test the null that the population means are the same, against the alternative that they are different? ...

Two-sample t-test

- Do two different populations have the same mean?
- Test-statistic:

$$T = \frac{\bar{y} - \bar{x}}{SE}$$

- Where
  - $\bar{y}$ and $\bar{x}$ are the average of the observations in both populations
  - SE is the standard error for the difference in means
  - There are a few different ways to compute SE, depending on assumptions
  - One possibility: $SE = \sqrt{S_y^2/n + S_x^2/m}$
- If H is true (no difference in mean), T follows a *t*-distribution with n + m - 2 degrees of freedom (where n and m are the number of observations in each sample)

Paired *t*-test

- In the two-sample *t*-test, we compared two samples of *unrelated* data points
- If the data between the two sample is *paired*, that is, each point x_i in the first sample correspond to a point y_i in the second sample, can do *paired t-test*
- *More powerful* than the two-sample *t*-test because we provide more information (the pairing) to the test

Paired t-test - calculation

- *Example:* 10 patients on diet; measure weight before they start the diet and after 6 months
- Does the group show a significant weight difference?
- In practice: calculate the difference $x_i - y_i$ between the two measurements, use one-sample t-test to test if this difference is significantly different from 0
- 'Different from': bigger, smaller, not equal (depending on the direction of interest)

Difference between two-sample and two-tailed tests

- A *two-sample test* is a hypothesis test for answering questions about means for *two different populations*
- Data are collected from two random samples of independent observations
- A *two-sided test* is a hypothesis test in which the values for rejecting the NULL are in *both tails of the probability distribution*
- The choice between a one-sided test and a two-sided test is determined by the purpose of the investigation or prior information

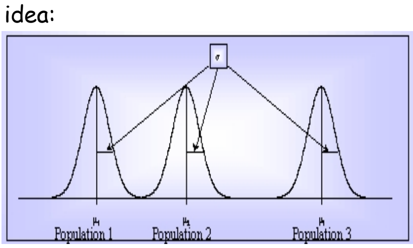
Even more trees

- You guessed it! Now say we are also interested in trees at 975 meters as well
- Want to make a three-way comparison
- Have a random sample (size 10 again) and find the mean is 2.5 and the SD is 1
- How might we test the null that all three means are the same, against the alternative that they are different?

(BREAK)

ANOVA

- Stands for ANalysis Of VAriance
- But it's a test of differences in *means*
- The idea:



The Observations y_{ij}

Treatment group

i = 1	i = 2	…	i = k
y_{11}	y_{21}	…	$y_{k,1}$
y_{12}	y_{22}	…	$y_{k,2}$
…	…	…	…
$y_{1, n1}$	$y_{2, n2}$	…	$y_{k, nk}$

means: m_1 m_2 … m_k

The ANOVA table

- The analysis is usually laid out in a table
- TS is an *F*-statistic with df k-1, n-k
- For a one-way layout (where the response is assumed to vary according to grouping on one factor):

Source	df	SS	MS	F	p-val
Treatment	k-1	$\Sigma(m_i-m)^2$	SST/(k-1)	MST/MSE	*
Error	n-k	$\Sigma(y_{ij}-m_i)^2$	SSE/(n-k)		
Total	n-1	$\Sigma(y_{ij}-m)^2$			

m = overall mean, m_i = mean within group *i*

Example (I)

- Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress
- An exercise physiologist measured the resting (unstressed) blood concentration of HBE in three groups of men aged 25-30: 16 who had just entered a physical fitness program, 11 who had been jogging regularly for some time, and 9 non-exercising people
- Want to compare HBE levels for these conditions

Example (II)

- The HBE levels (pg/ml) are summarized in the table
- Estimate the average increase in resting HBE level of the non-exercising group over
  - the joggers
  - the program participants

	Program	Joggers	Non-exercisers
Mean	38.7	35.7	42.5
SD	16.0	13.4	12.8
n	16	11	9

Example (III)

- When we want to evaluate a study, need to consider *several aspects:*
- Is the study observational or experimental*??*
- Is there likely to be a placebo effect in this study*??*
- Are there any possible confounders*??*
- State in words the null hypothesis for testing equality of all three means *…*

Example (IV)

- Calculations based on the raw data yielded a sum of squares (SS) Between Groups = 240.7 and SS Total = 7128.3
- Complete the ANOVA table
- What do you conclude*??*

Source	*df*	*SS*	*MS*	*F*	*p*
Treatment		240.7			
Error					
Total		7128.3			

Assumptions

- Have *random samples* from each separate population
- The *variance is the same* in each treatment group
- The individual population distributions are *normal* (or: the samples are *sufficiently large* that the CLT holds for each sample mean)

What does it mean when we reject H?

- The null hypothesis H is a *joint* one: that *all* population means are equal
- When we reject the null, that does *NOT* mean that the means are all different!
- It means that *at least one* is different
- To find out *which* is different, can do *post hoc* testing (pairwise *t*-tests, for example)

Additional aspects

- Why not start off doing separate (z or t) tests for each pair of samples? *...*
- Testing the assumptions
- *Which* mean(s) is/are not equal
  - can do *post hoc* testing (pairwise *t*-tests, for example)
- Multiple comparisons (multiple testing)
- 'Data snooping'

Factorial crossing

- Compare 2 (or more) sets of conditions in the *same experiment*
- Designs with factorial treatment structure allow you to measure *interaction* between two (or more) sets of conditions that influence the response – you will look at this in more detail during the exercises today
- Factorial designs may be either observational or experimental

3 types of 2-factor factorial designs

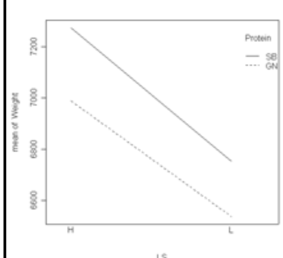
- 2 experimental factors – you *randomize* treatments to each unit
- 2 observational factors – you *cross-classify* your populations into groups and get a sample from each population
- 1 experimental and 1 observational factor – you *get a sample* of units from each population, *then use randomization* to assign levels of the experimental factor (treatments), separately within each sample

Interaction

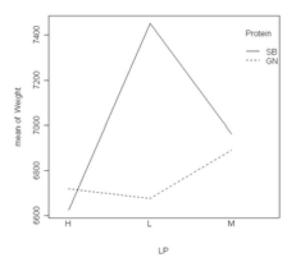
- Interaction is very common (and very important) in science
- Interaction is a *difference of differences*
- Interaction is present if the effect of one factor *is different* for different levels of the other factor
- *Main effects can be difficult to interpret in the presence of interaction*, because the effect of one factor depends on the level of the other factor

Interaction plot

no interaction



interaction



Testing a population mean

- We have already learned how to test the
  mean
  of a population for a variable with a
  normal distribution
  when the sample size is
  small
  and the population
  SD is unknown
- *What test is this??*

Assumption of normality

- *t*-test (and ANOVA) developed for samples that have *normally distributed* values
- These are examples of *parametric tests* – a (parametric) form of the distribution is assumed (here, a normal distribution)
- The tests are fairly robust against departures from normality if the sample size is not too small
- *BUT* if the values are extremely non-normal, it might be better to use a procedure which does not make this assumption

Nonparametric hypothesis tests

- *Nonparametric* (or *distribution-free*) hypothesis tests do not make assumptions about the *form* of the distribution of the data values (*e.g.* do not assume normality of the observations)
- These tests are usually based on the *ranks* of the values, rather than the actual values themselves
- There are nonparametric analogues of many parametric test procedures

One-sample Wilcoxon test

- Nonparametric alternative to the *t*-test
- Tests value of the center of a distribution
- Based on sum of the (positive or negative) *ranks* of the differences between observed and expected center
- Test statistic corresponds to selecting each number from 1 to n with probability $\frac{1}{2}$ and calculating the sum
- In R: `wilcox.test()`

Two-sample Wilcoxon test

- Nonparametric alternative to the 2-sample *t*-test
- Tests for differences in location (center) of 2 distributions
- Based on replacing the data values by their ranks (without regard to grouping) and calculating the sum of the ranks in a group
- Corresponds to sampling n_1 values without replacement from 1 to $n_1 + n_2$
- In R: `wilcox.test()`

Matched-pairs Wilcoxon

- Nonparametric alternative to the paired *t*-test
- Analogous to paired *t*-test, same as one-sample Wilcoxon but on the *differences* between paired values
- In R: `wilcox.test()`

ANOVA and the Kruskal-Wallis test

- Nonparametric alternative to one-way ANOVA
- Mechanics similar to 2-sample Wilcoxon test
- Based on between group sum of squares calculated from the average ranks
- In R: **`kruskal.test()`**

Issues in nonparametric testing

- Some (mistakenly) assume that using a nonparametric test means that you don't make any assumptions at all
- THIS IS **NOT** TRUE!!
- In fact, there is really only one assumption that you are relaxing, and that is of the *form* that the distribution of sample values takes
- A major reason that nonparametric tests are avoided if possible is their relative *lack of power* compared to (appropriate) parametric tests