Hypothesis Testing
Notes updated from Philip Sabes and Loren Frank, UCSF
Suggested reading: Book chapters 11,12,15

1. What is a "hypothesis" and how do you test it?

a. Hypotheses in the colloquial sense

b. A formal hypothesis:

i. A contrast between two alternatives:

1. H0 – "Null hypothesis"
2. H1 – "Alternative hypothesis", can be multiple H1's for a given H0
3. Example: flip a coin N times and heads each time
    a. What's H0, H1?
    b. See below for Hypothesis test

c. Hypothesis Testing

i. Formulate H0, H1

ii. Formulate quantitative relationship to **observables** (a **"model"**)

1. What is a "model" for the example above
2. Here is where the **assumptions** come in

iii. Choose test. This typically entails:

1. A **statistic**, i.e. a scalar function of the data
2. A distribution of the statistic under H0

iv. Two results are possible:

1. **reject** H0 in favor of H1
2. **do not reject** H0
3. Significance level
    a. α = Type I / "false positive", P(reject H0|H0)
    b. Reject H0 is p(statistic|H0)< α
    c. $1-\alpha$ "confidence level"
4. Can never "prove" H0 or H1. Why?

v. Power

1. β = Type II errors/"false negative", P(accept H0|H1)

2. Power = 1-β, i.e. "hit rate", P(reject H0|H1)

2. The Binomial Test

a. **The problem:** Say you are playing a coin-toss game in which heads beats tails. In her first N tosses, your opponent manages get M heads, where M is noticeably bigger than N/2. Is your opponent cheating, i.e. is her coin biased?

b. The hypotheses

i. H0 – the coin is unbiased

ii. H1 – the coin is biased towards heads
Note it's not the case that H1 == ~H0 (i.e., "H1 is not equal to 'not H0'", using Matlab notation).

c. The model (assumptions)

i. Say P(heads)=p, P(tails)=1-p for a single coin toss

ii. What is the probability of M heads given N coin tosses?
Binomial distribution:

$$P(M \mid N, p) = \binom{N}{M} p^M (1-p)^{N-M}$$

d. The statistic and test

i. Statistic is just the number of heads, M

ii. Reject H0 if the probability of getting M or greater heads is less than α under the null hypothesis, i.e.

$$\text{Reject H0 if } \sum_{k=M}^{N} P(k \mid N, p = 0.5) < \alpha$$

iii. In Matlab, can use binocdf (see problem set).

e. **One-tailed** vs **two-tailed** tests

i. In the example above, we specifically wanted to test whether the coin was towards heads. So we were in interested in the probability of getting M or greater heads, given H0. This is a "one-tailed" test.

ii. We may want to ask the more general question of whether the coin is biased in either direction. In this case, we want to consider the probability of getting M or greater heads **OR** M or greater tails, since either would have appeared equally "suspicious". The test is now:

$$\text{Reject H0 if } \sum_{k=0}^{N-M} P(k \mid N, p=0.5) + \sum_{k=M}^{N} P(k \mid N, p=0.5) < \alpha .$$

This is "two-tailed" test.

iii. If should be clear that the sums in the two-tailed test are larger than in the one-tailed test, so an M closer to 0 or N is needed to obtain significance at the same level. This makes sense, since it easier to "look suspicious" by chance when you it can happen on either tail.

f. Examples – see problem set.

g. Other comments:

i. When N is large, you can use an approximation (Chi-squared test). Not as important with fast computers.

ii. Can test for other H0's, e.g. p=1/6 to test whether a die is throwing too many 6's.

iii. Multinomial test based, on multinomial distribution, when there are more than two possible outcomes. The null hypothesis in this case takes the form of a vector of probabilities

$$\text{H0: } (p_1, \ldots, p_k)$$

and the probablility distribution is

$$p(\{M_i\} \mid N, \{p_i\}) = N! \prod_{i=1}^{k} \frac{p^{M_i}}{M_i!}$$

3. Testing the sample mean: Student's t-test

a. **The problem:**

i. Say you collect measure the membrane potential of N neurons in some altered physiological setting. You want to know if the membrane potential is different from some known baseline value. The problem is that these measurements are noisy, and you are not sure the difference you see is "real" or just noise.

ii. More generally: you have a set of N repeated measurements x_i, i=[1 N], of some variable of interest. You want to know if the true mean of the distribution from which the x_i are drawn, μ=E(x), s different from some predetermined expectation, μ_0.

b. The hypotheses

i. H0: $\mu = \mu_0$

ii. H1: $\mu \neq \mu_0$

c. The model (assumptions) Check: Is it a gaussian distribution?

i) One sample – comparison of mean to constant. What is the probability that the estimate of the mean is equal to the constant.

i. Assume that the x_i are "i.i.d. Normal random variables": identically, independently distributed. The pdf is then:

$$f(x_i \mid \mu,\sigma) = N(\mu,\sigma)$$

$$f(\{x_i\} \mid \mu,\sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

d. The statistic and test

i. We use the following statistic:

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \qquad \text{OR} \qquad \text{Use standard error of the mean.}$$

where $\bar{x}$ is the sample mean and *s* is the sample standard deviation.

ii. WHY?!? Because a clever statistician (name Gosset, pen name Student) figured out that (or something like it) would have a calculable distribution (more specifically, a calculable cdf). Student's t-distribution is derived from two previously known distributions:

1. t = z/s. where
2. z is "standard normal", N(0,1)
3. s is Chi-squared with ν=N-1 degrees of freedom

iii. Digression on "degrees of freedom": number of free parameters

1. "number of values in the final calculation of a statistic that are free to vary" http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)
2. When the sample standard deviation is computed, the mean is subtracted first, removing one degree of freedom, hence N-1.

iv. The distribution of *t* depends only on the degrees of freedom, ν.

v. The test: Reject H0 if

i) 1-F(t|ν) < $\alpha/2$ ($\bar{x}$ is bigger than expected)
ii) F(t|ν) < $\alpha/2$ ($\bar{x}$ is smaller than expected)

Why $\alpha/2$ and not α? This is a **two-tailed test** (H1 is $\mu \neq \mu_0$).

vi. If instead we want to test a specific direction of deviation, e.g. we expect that our physiological manipulation will increase the membrane potential and we test for that, then we want a **one-tailed test,** in this case H1 is $\mu > \mu_0$. This one-tailed test is:

Reject H0 if 1-F(t|ν) < α

vii. In Matlab, can use *tcdf* or even just *ttest*:

[H,P,CI,STATS] = ttest(X,M,ALPHA,TAIL)

1. CI contains the confidence interval for the true mean:

$$-t_{\alpha,\nu} \le \frac{\overline{X} - \mu}{s_{\overline{X}} / \sqrt{n}} \le t_{\alpha,\nu}$$

alpha = 0.05
95% confidence bound

$$\overline{X} - \frac{s_{\overline{X}} t_{\alpha,\nu}}{\sqrt{n}} \le \mu \le \overline{X} + \frac{s_{\overline{X}} t_{\alpha,\nu}}{\sqrt{n}}$$

2. STATS contains t, ν, and s.

4. Comparing two sample means: the two-sample t-test

a. The problem:

i. Consider the same example as above (Section 3a), except that this time you bothered to do the proper controls. That is, you measured the membrane potential of a bunch of cells in the "baseline" condition and then you measured the membrane potential of a bunch of cells in some altered/experimental condition. You want to know if the membrane potential is different in these two groups. Again, the problem is that these measurements are noisy, and you are not sure the difference you see is "real" or just noise.

ii. More generally: you have two sets of N repeated measurements x_i, y_i. You want to know if the true means of the distributions from which the x_i and y_i are drawn, μ_x=E(x) and μ_y=E(y), are the same or different.

b. The hypotheses

i. H0: $\mu_x = \mu_y$

ii. H1: $\mu_x \neq \mu_y$

c. The model (assumptions)

i. Assume that the x_i , y_i are each "i.i.d. Normal" samples. For now, we assume that the sample sizes are the same (N_x=N_y=N) and so are the standard deviations (σ_x=σ_y=σ). The pdf of the data are then:

$$f(\{x_i, y_i\} \mid \mu_x, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_x)^2}{2\sigma^2}\right) \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_y)^2}{2\sigma^2}\right)$$

d. The statistic and test

i. We use the following statistic:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{1}{n}\left(s_x^2 + s_y^2\right)}}$$

where $\bar{x}$, $\bar{y}$ are the two sample means and s_x, s_y *are* the two sample standard deviations.

ii. This also follows Student's t-distribution, with ν=XXX [can you figure it out?]

iii. The **two-tailed** test:

$$\mathrm{Var}\left[\overline{X_2} - \overline{X_1}\right] = \mathrm{Var}\left[\overline{X_2}\right] + \mathrm{Var}\left[\overline{X_1}\right] = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

Reject H0 if

i) 1-F(t|ν) < α/2 ($\bar{x}$ is bigger than expected)

ii) F(t|ν) < α/2 ($\bar{x}$ is smaller than expected)

iv. **One-tailed** test ("right"): Reject H0 if 1-F(t|α <)ν

v. In Matlab, can use *tcdf* or even just *ttest2 (if variances are equal)*.

vi. If variances are not equal, look up t-valus using tcdf for reduced degrees of freedom

e. Sample sizes and variances. We assumed that the sample sizes are the same and that the variances are equal. These assumptions are "easily" relaxed, but of course you need a different statistic. See http://en.wikipedia.org/wiki/T-test and 'help ttest2'.

f. **Paired vs. unpaired t-test.** The development above is for "independent sample" or "unpaired" t-test. Why? Because we collected our baseline and experiments measurements on different neurons.
If we had collected baseline and experimental measurements on **each** neuron, then we would have "dependent sample" or "paired t-test". In practice, however, the paired t-test is just a one sample test with random variables $d_i=x_i-y_i$ and $\mu_0=0$. So you can use Matlab's ttest.

5. Multiple Comparisons: ANOVA
One-way ANOVA
Multi-way ANOVA

6. What about those assumptions?

a. "Linear gaussian"

b. Sign Tests in place of t-test

i. Look at distribution of sign of differences

ii. Under H0, p(d<0)=p(d>0)=0.5

iii. Reduces to a Binomial test

c. Other non-parametric tests, e.g. Rank-order tests
http://en.wikipedia.org/wiki/Non-parametric_statistics

d. Why does it matter?

7. Permutation tests

a. A general, powerful approach to hypothesis testing:

   i. Can test **any** statistic from any dataset! Don't need to be a "Fisher" or a "Student".

   ii. Very few assumption (e.g. don't have to assume Linear/Gaussian)

   iii. BUT…. Much more computationally demanding

b. The big idea: jumble the data so that

   i. Our statistic of interest should not change under H0

   ii. We "break" H1 so the jumbled data **really do** obey H0, even if the original data obeyed H1

c. Example: two-sample t-test

   i. We have two unpaired datasets x_i, i=[1 N], y_i, i=[1 N]

   ii. H0: $\mu_x = \mu_y$, H1: $\mu_x \neq \mu_y$

   iii. We are interested in the difference between the sample means:

$$d = x - y$$

   iv. Under H0, E(d)=0. Under H1, E(d)≠0.

   v. How do "jumble" (permute) our data?

      1. Under H0, d should be zero.

      2. In fact, under H0 it doesn't really matter which data are "x" and which are "y", since they all have the same mean.

      3. So we can put all 2N (or Nx + Ny) data into a big bucket, "jumble" them, and then pull out Nx and Ny at random and call them the "permuted" $\{x_i\}$ and $\{y_i\}$ datasets. We can then compute a new difference, $d_{perm.}$

      4. $E(d_{perm})=0$ under H0 or H1, i.e. the permutation "breaks" H1.

      5. More importantly, the expected distribution of d_{perm} is the same as that of d under H0.

   vi. The test:

      1. Compute $d = \bar{x} - \bar{y}$ from the original data.
      2. Repeat the following R times (R = 1000?)

         a. Permute the x/y identities of the dataset to get a new permuted dataset

b. Compute the r^{th} value of $d_{perm,r}$ using the permuted data

3. This yields a set of R values $\{ d_{perm,r}\}$
4. Reject H0 if the original d is greater than the $\alpha/2$th or less than the $1-\alpha/2$th percentile value of the $\{ d_{perm,r}\}$.

8. Power Analysis

a. Type I error: incorrect rejection of H0, "false positive", p(reject H0 | H0) = α

b. Type II error: incorrect acceptance of H0, "false negatgive", p(~reject H0 | H1) = β

c. Power: rate of rejecting H0 given H1, p(reject H0 | H1) = $1 - \beta$.

d. Typically, you pick an effect size and then determine how many samples you need to detect that effect with a given significance level (Type II error rate), α.

e. Example: t-test (see Problem Set)

9. References:

a. Matlab Help Browser. (Contents):
Statistics Toolbox : Probability Distributions : Supported Distributions
This contains some helpful tables.

b. Wikipedia!

t-tests (additional info)

i) One sample – comparison of mean to constant
   (1) What is the probability that the estimate of the mean is equal to the constant.
      1. $t = \dfrac{\overline{X} - c}{s_{\overline{X}}}$ where c is the constant and
         $s_{\overline{X}}$ is the standard error of the mean.
      2. We then look up the value of t for the number of degrees of freedom.
   (2) Matlab: [H,P,CI,STATS] = TTEST(X,M,ALPHA,TAIL)
   (3) Confidence limits
      (a) Get values for t representing middle 95% of distribution
         i. $-t_{0.05,v} \le \dfrac{\overline{X} - c}{s_{\overline{X}}} \le t_{0.05,v}$
         95% confidence bound: $\overline{X} - s_{\overline{X}} t_{0.05,v} \le c \le \overline{X} + s_{\overline{X}} t_{0.05,v}$

ii) Two sample

$$t' = \frac{\overline{X_2} - \overline{X_1}}{s_{\overline{X_2} - \overline{X_1}}}$$ where $s_{\overline{X_2} - \overline{X_1}}$ is the standard error of the difference.

$$\mathrm{Var}\left[\overline{X_2} - \overline{X_1}\right] = \mathrm{Var}\left[\overline{X_2}\right] + \mathrm{Var}\left[\overline{X_1}\right] = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

$$s_{\overline{X_2} - \overline{X_1}} = \sqrt{\mathrm{Var}\left[\overline{X_2} - \overline{X_1}\right]} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

iii) If the variances of the two populations are equal, you can use a normal two sample t-test (TTEST2 in Matlab).
iv) If the variances are not equal, you should use a modified version where the number of degrees of freedom is reduced: $\mathrm{df} = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
v) Look up t value (Matlab: TCDF)

e) Testing for equality of variance
   i) $F = \arg\max\left(\dfrac{s_1^2}{s_2^2}, \dfrac{s_2^2}{s_1^2}\right)$
   ii) Look up value of F CDF with degrees of freedom from distributions 1 and 2 (or 2 and 1).

f) ANOVA
   i) For > 2 groups
   ii) Assumptions
      (1) Normality

(2) Homogeneity of variance
(3) Independence of observations

iii) Important to keep family-wise error rate down
(1) Keep probability that results occurred by chance at a reasonable level.

iv) Example
(1) Suppose we have three groups with stdevs of $\sigma_1^2, \sigma_2^2,$ and σ_3^2
(a) We measure the means of these groups to be $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3$
(b) We measure the total mean as $\bar{X}_{total} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3}$ and the variances of the mean as $s_{\bar{X}}^2 = \frac{\sum_k \left(\bar{X}_k - \bar{X}_{total}\right)^2}{k-1}$ where k is the number of groups.
(c) Note the use of k-1 instead of k.
(d) If the three groups come from the same distribution, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_e^2$.
(e) One estimate of σ_e^2 is the mean of the measured variances, or $\sigma_e^2 = (s_1^2 + s_3^2 + s_3^2)/3$.
(f) Another estimate $\frac{\sigma_e^2}{n} = s_{\bar{X}}^2$, so $\sigma_e^2 = n s_{\bar{X}}^2$ where n is the number of measurements in each group
We can then compare the two estimates of variance to determine if they are equal.

v) One way ANOVA
(1) Matlab: [P,ANOVATAB,STATS] = ANOVA1(X,GROUP,DISPLAYOPT)
(a) X is a list of all of the variables
(b) GROUP is the same size as X and all values of X whose group is the same are grouped together.
(c) STATS is used for post-hoc comparisions

vi) Multiple groups
[P,T,STATS,TERMS]=ANOVAN(Y,GROUP,MODEL,SSTYPE,GNAME,DI S PLAYOPT) performs anova on the vector Y grouped by entries in the cell array GROUP.
GROUP must be a cell array, with each cell containing a grouping variable. The grouping variable can be a numeric vector, a character matrix, or a single-column cell array of strings. Each grouping variable must have the same number of items as Y.
MODEL is an indication of the model to be used:
'linear' to use only main effects of all factors (default)
'interaction' for main effects plus two-factor interactions
'full' to include interactions of all levels
an integer representing the maximum interaction order, for example 3 means main effects plus two- and three-factor interactions
a vector V of integers, with each element describing a term, so

the Ith term includes the Jth grouping variable if
BITGET(V(I),J)==1
SSTYPE is 1, 2, or 3 for the type of sum of squares to use (default 3).
GNAMES is a character matrix of names for the grouping variables
(default has rows 'X1','X2',...).
DISPLAYOPT is 'on' (default) to display table, 'off' to omit display
P is a returned vector of p-values, one for each term.

g) Multiple comparisons (after anova)
   i) Often desirable to identify significant differences between individual groups without increasing family-wise error rate
   ii) Types
      (1) Tukey-Kramer
         (a) Order means, compare groups
      (2) Scheffé test
         (a) Compare any arithmetic combination of means

h) Power analyses
   i) Compute N's necessary given hypothesized effect size

Descriptive statistics – non-gaussian data

a) Comparison of proportions
   i) Z-test for proportions
      (1) Compare observed proportion to hypothesized proportion
      (2) Use binomial probability model
         (a) Given expected proportion, we can compute standard error
         (b) For binomial, standard error of mean is $\sqrt{\frac{p(1-p)}{n}}$
         (c) Compute $z = \frac{p_{observed} - p_{predicted}}{\sqrt{\frac{p_{predicted}(1-p_{predicted})}{n}}}$, z should be from a standard normal distribution, so we can then look up the probability of the observed data.
   ii) Z-test for two proportions
      (1) Compare observed proportions
      (2) Use binomial probability model
         (a) Null hypothesis: both sets of data came from the same distribution with underlying probability of success *p*.
         (b) If so, then our best estimate is $p_{total} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$
         (c) For binomial, standard error of mean is $\sqrt{\frac{p(1-p)}{n}}$

(d) Compute $z = \frac{p_{observed} - p_{predicted}}{\sqrt{\frac{p_{total}(1-p_{total})}{n}}}$, z should be from a standard normal distribution, so we can then look up the probability of the observed data.

b) Comparisons between two groups

i) **Sign test**

(1) Based on binomial distribution

(2) Used for paired measurements, ask whether one group is more often larger than the other

(3) E.g. two groups, group A > group B in 2 of 10 cases.

(a) Compute

$$\Pr\left(X \le 2 \text{ or } X \ge 8\right) = \Pr(X=0) + \Pr(X=1) + \Pr(X=2) + \Pr(X=8) + \Pr(X=9) + \Pr(X=10)$$

(b) In this case it is about 0.1

ii) **Wilcoxon signed-rank test**

(1) Used for paired samples

(2) Tests for differences in median between groups

(3) Matlab: [P,H,STATS] = SIGNRANK(X,Y,ALPHA)

iii) **Wilcoxon Rank-Sum test**

(1) Comparison of ranks of data

(2) Combine data from two groups into a single, sorted list

(a) Record the ranks (position in total list) for one group

(b) Matlab: [P, H, STATS] = RANKSUM(X,Y,ALPHA)

iv) **Kolmogorov-Smirnov test**

(1) Test distribution of data against either

(a) Known CDF (Matlab: KSTEST)

(b) Another CDF (Matlab: KSTEST2)

(c) Compares CDFS of functions and finds point of maximum distance D

c) **More than 2 groups**

i) **Non-parametric ANOVA**

(1) Kruskal-Wallis – ANOVA based on ranks of data

(2) Friedman – Two way ANOVA based on ranks of data

(3) Both can be used with MULTCOMPARE

ii) **Chi-squared**

(1) Build contingency table

	Mutant	Control
Male	5	10
Female	10	5

(2) Using a null hypothesis, compute number expected as the marginal total of each row times the marginal total of each column divided by the total number

	Mutant	Control	Totals
Male	5	10	15
Female	10	5	15

Totals	15	15	30

Expected	Mutant	Control
Male	7.5	7.5
Female	7.5	7.5

(3) Compute chi-square statistic

$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}}$ where f_{ij} is the number observed

and $\hat{f}_{ij}$ is the number expected.

(4) Look up in table (Matlab: chi2cdf), degrees of freedom = (n-1)(m-1)

Circular Statistics

a) Data distributed around a circle
   i) Movement direction
   ii) Spiking in relation to a rhythm
   iii) Measures
      (1) Mean angle of $a_1,...a_n$

$$X = \frac{\sum_{i=1}^{n}\cos(a_i)}{n}, Y = \frac{\sum_{i=1}^{n}\sin(a_i)}{n}]$$

$$\bar{a} = \tan^{-1}\left(Y/X\right)$$

      (2) Angular dispersion

$r = \sqrt{X^2 + Y^2}, 0 \le r \le 1$

      (3) Circular variance

$S^2 = 1 - r$

Angular variance

$s^2 = 2(1-r)$

b) Hypothesis testing
   i) Watson U test
      (1) Compares CDFs of angular distributions
   ii) Watson Williams test
      (1) Compares the means of two angular distrubtions
   iii) Not included in Matlab, use versions on course website.