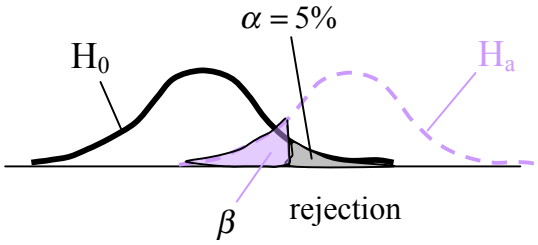
Hypothesis testing (Chapter 5 of Wilks)

Introduction:

Consider Table A.3 of Wilks, with T, SLP and pp in Guayaquil, Ecuador, for June over 20 years, five of which are El Niño years. It is obvious by inspection that it seems to rain more in an El Niño year. It also seems like in those years the temperature tends to be higher and the pressure lower, but how do we know that it is not just sampling? Hypothesis testing allows us to state "during the El Niño years the pressure is below normal" with a confidence interval of, for example, 95%, i.e., the probability of having obtained this experimental result by sampling fluctuations is less than 5%, or one in 20. One does that by creating a probability distribution corresponding to the "null hypothesis", i.e., that El Niño is not related to the surface pressure. Then we estimate the probability that we observe as many cases of low pressure for El Niño as we actually observed, and if it is less than 5%, we reject the null hypothesis.

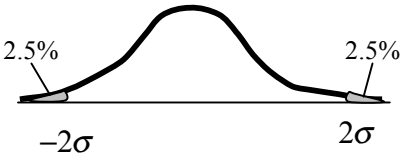
Parametric testing (theoretical): probabilities of a null hypothesis derived from a theoretical PDF.
Non-parametric testing: No PDF assumed. Data is resampled to derive probability of null hypothesis from the sampled data itself.
Sample statistics: μ and σ are estimated by $\bar{x}$, s: they can fluctuate due to sampling.
Hypothesis testing, steps:

1) Choose the test statistics for a given data, e.g., mean, trend, and a test level α, e.g., 5%.
2) Define null hypothesis, H_0: e.g., two samples belong to the same population, or there is no trend. Usually we would like to reject it.
3) Define alternative hypothesis H_a: that H_0 is not true. Can be one-sided (there is a warming trend) or two sided (the two samples belong to different populations).
4) **Consider or create the null distribution: assume H_0 is true, and obtain statistics for H_0.**
5) Compare the test statistics to the null distribution. Obtain the probability p of the test statistic to be observed in the null distribution. If the p-value (probability of finding this sample mean or trend within the null distribution) is less than the test level, $p < \alpha$, then the null hypothesis is rejected.



The test can give wrong results due to sampling:
Type 1 error: $p < \alpha$ but H_0, the null hypothesis is true: H_a, the alternative hypothesis is accepted but it is not true. Wrong rejection of H_0 because the sample is biased away from H_0!
Type 2 error: H_o is not rejected, but H_a is true (area β). Wrong rejection of H_a.

One-sided versus two-sided test:



$$P\{|\bar{x} - \mu| > 2\sigma\} = 1.96\sigma \simeq 2\sigma$$

The alternative hypothesis determines whether it is a one or two "tailed" test: H_a=not null hypothesis➔2-tail test; H_a: $\mu > \mu_0$, one tail.

Example of parametric test: Assume that on a given day P(rain)=0.1= μ.
It rains 2 days out of 5: is this sample significantly different from the assumed population? Null hypothesis: it belongs to the population. Alternative hypothesis: it rained too much: the probability of having 2 (or more) days of rain out of 5 is too low for the sample to belong to the population.

The null population has a Binomial distribution: $P(X \geq 2) = \sum_{x=2}^{5} \binom{5}{x} 0.1^x 0.9^{5-x}$

This can be approximated with the Poisson $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}, \quad \mu = 0.1$

$$P(X=0)=\frac{1*e^{-0.1}}{1}=0.905$$
$$P(X=1)=\frac{0.1*e^{-0.1}}{1}=0.090$$
$$P(X=2)=\frac{0.01*e^{-0.1}}{2}\simeq 0.005$$
$$P(X=3,4,5)\simeq 0$$

Therefore the sample is "different" with a 1% level of significance.

One sample t-test (parametric): Compare a sample mean with a population

$t_\nu=\dfrac{\bar{x}-\mu_0}{\left(\hat{\mathrm{var}}(\bar{x})\right)^{1/2}}$; $\hat{\mathrm{var}}(\bar{x})=\dfrac{s^2}{n}$. Here $\nu=n-1$ is the number of degrees of freedom (one was used to compute $\bar{x}$).

Test of the difference between two samples (assuming they are independent, not paired):

$t_\nu\simeq z=\dfrac{(\bar{x}_1-\bar{x}_2)-E(\bar{x}_1-\bar{x}_2)}{\left[\hat{\mathrm{var}}(\bar{x}_1-\bar{x}_2)\right]^{1/2}}=\dfrac{(\bar{x}_1-\bar{x}_2)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}$ with $\nu=n_1+n_2-1$ d.o.f. We have used the null hypothesis to assume $E(\bar{x}_1-\bar{x}_2)=0$.

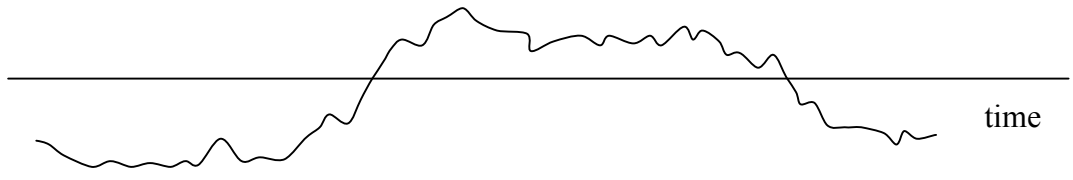
If the two samples are paired

$$\hat{\mathrm{var}}(\bar{x}_1-\bar{x}_2)=\hat{\mathrm{var}}(\bar{x}_1)+\hat{\mathrm{var}}(\bar{x}_2)-2\hat{\mathrm{cov}}(\bar{x}_1,\bar{x}_2)=\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}-2\rho_{1,2}\sqrt{\frac{s_1^2}{n_1}\frac{s_2^2}{n_2}}$$ with $n_1=n_2$.

$z=\dfrac{\bar{x}_1-\bar{x}_2}{\sqrt{\left(s_1^2+s_2^2-2\rho_{12}s_1s_2\right)/n}}$. The correlation increases the significance of the difference between pairs if $\bar{x}_1\neq\bar{x}_2$.

Tests for data with persistence

Example of a persistent time series with long time mean=0. Short time averages have an error larger than $\sqrt{s^2/n}$ because the n measurements are not independent



Because of persistence the observations are not "independent". Time averages will tend to drift away from the long-term mean (persistent anomalies). Therefore the number of degrees of freedom (independent observations) is smaller. Estimated as

$\hat{\text{var}}(\bar{x}) \simeq \frac{s^2}{n'} = \frac{s^2}{n}\left(\frac{1+\rho_1}{1-\rho_1}\right)$ variance inflation.

n': the number of effectively independent samples.
ρ_1: 1-day lag correlation.

Summary of parametric hypothesis typical tests:

Here we review most cases of hypothesis that appear in practical applications and the corresponding test that is applied.

Z: standard normal (Gauss) distribution, used if you know the variance of the population
T_{n-1}: student t distribution with n-1 d.o.f., used if you estimate the standard deviation from the sample
α: level of significance (e.g., 5%=0.05)
H_a: the alternative hypothesis that determines whether it is a one-tailed or two-tailed problem.

1) Test whether a sample with mean $\bar{X}$ belongs to a population with mean μ_0, assuming the sample has the same (known) standard deviation σ (two-tailed problem).

$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2 / n}}$; find the critical value $z_{\alpha/2}$ such that $P\{|Z| \le z_{\alpha/2}\} = 1-\alpha$. If α=5%, then $z_{\alpha/2} = 1.96 \approx 2$

In other words, if $|Z| = \left|\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2 / n}}\right| > 2$ we reject that the sample mean $\bar{X}$ belongs to a population with mean μ_0.

Probability that a result was obtained by chance: "p-value"

If $P\{|Z| \le z_{\alpha/2}\} = 1-\alpha$ then $P\{|Z| \ge z_{\alpha/2}\} = \alpha$. So, $1-\alpha$ is the level of significance (e.g., 95%) and α is the probability of obtaining this result

by chance (e.g., $\alpha = 0.05$ or 5%). If $P\{|Z| > z_{\alpha/2}\}$ then the probability of getting this value of |Z| by chance (the "p-value") is $p < \alpha$ (see table below).

Level of significance	Critical value of |Z|	p-value
0.80	1.28	p<0.20
0.90	1.64	p<0.10
0.95	1.96	p<0.05
0.99	2.58	p<0.01
0.999	3.29	p<0.001
0.9999	3.89	p<0.0001
0.999999	4.89	p<0.000001
0.99999999	6.11	p<0.00000001

2) Test whether a sample with mean $\bar{X}$ belongs to a population with mean μ_0, but estimating the unknown standard deviation *s* from the sample (two-tailed problem).

$T_{n-1} = \dfrac{\bar{X} - \mu_0}{\sqrt{s^2 / n}}$ $\quad s^2 = \dfrac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$; find the critical value $t_{\alpha/2,n-1}$ such that $P\{|T_{n-1}| \le t_{\alpha/2,n-1}\} = 1-\alpha$. If α =5%, then for n-1=10, $t_{\alpha/2,10} = 2.2$

3) Test the equality of means of two samples, assuming the s.d. are known

$Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$. Then look for $P\{|Z| \le z_{\alpha/2}\} = 1-\alpha$; with α =5%, then

$z_{\alpha/2} = 1.96$

4) Test whether two samples belong to the same population: by far the most common test in practice

$$T_{n_1+n_{2-2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 / \left(1/n_1 + 1/n_2\right)}},$$

where the "pooled variance" is $s_p^2 = \frac{(n_1-1)^2 s_1^2 + (n_2-1)^2 s_2^2}{n_1+n_2-2}$

Then check whether $P\left\{|\, T_{n_1+n_2-2} \,| \le t_{\alpha/2, n_1+n_2-2}\right\} = 1-\alpha$.

For n_1+n_2-2~10, $t_{\alpha/2,10} = 2.2$, so that if $T_{n_1+n_{2-2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 / \left(1/n_1 + 1/n_2\right)}} >\sim 2.2$ we reject the hypothesis that the two samples belong to the same population with a level of significance of 5%.

5) Paired tests of two time series: define $w_i = x_{1i} - x_{2i}, \quad i = 1,2,...n$

$T_{n-1} = \frac{\bar{W}}{\sqrt{s_w^2/n}}$; $P\left\{|\, T_{n-1} \,| \le t_{\alpha/2,n-1}\right\} = 1-\alpha$; If α =5%, then $t_{\alpha/2,10} = 2.2$

6) Test whether the variance of a sample $s^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$ is equal to the population variance σ_0^2. The variable $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$ has a chi-square distribution with n-1 d.o.f.

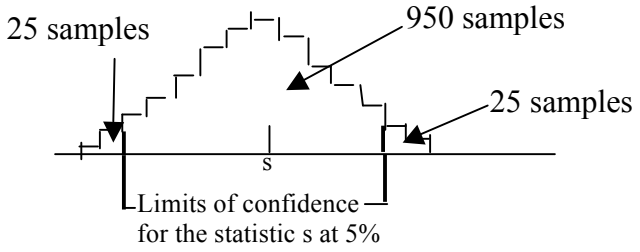
Example: n=11. From Table 3, if $3.247 \le \chi^2 \le 20.48$, values corresponding to $\alpha = 0.975$ and $\alpha = 0.025$ respectively, then the null hypothesis is accepted with a significance level of 5%.

7) To check whether the variances of two populations are equal, we use the F-test (Table 4): $F_{n_1-1,n_2-1} = \frac{s_{x_1}^2}{s_{x_2}^2}$ and compare with the value $F_{0.05,n_1-1,n_2-1}$ from Table 4.

Non-parametric tests based on resampling (bootstrapping)

Example 1: Determine the limits of confidence with which a statistic (e.g., mean $\bar{x}$, s.d. *s*. median, Inter Quartile Range IQR, trends, anything!) is estimated from a sample of size *n*.

We resample the batch of data by choosing a datum randomly and replacing it (without replacement we would only obtain the original *n* values). Easy way to sample: rank the data $x_i, i = 1,...n$. Pick random numbers *r* uniformly distributed between 0 and 1. If $j-1 < nr \le j$, pick the datum x_j. Create a large set of n samples (e.g., 1000 *n*-sized samples), and compute for each of them the statistic of interest. Plot a histogram, and the boundaries of 25 samples on both tails give the limits of confidence of the statistic *s*.

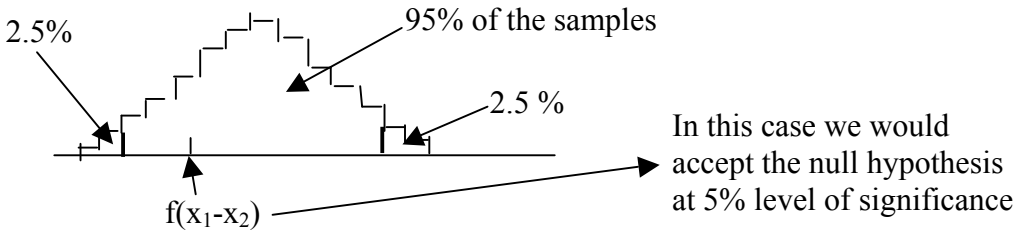


"bootstrapping"

Example 2: test whether two samples of size n_1 and n_2 belong to the same population. Null hypothesis: they are from the same population. So we create a "null population" by pooling the two samples, and create samples of size n_1, n_2 from the pooled n_1+n_2 sample. Since the number of possible choices increases fast with n_1, n_1+n_2, we have the luxury of creating samples without replacement (i.e., each combination n_1, n_2 is

picked only once). For example if $n_1 = n_2 = 5, \binom{10}{5} = 112$,

if $n_1 = n_2 = 10, \binom{20}{10} = 923,780$. Then we can test any statistic that compares the original two samples (e.g., $\bar{x}_1 - \bar{x}_2, |\bar{x}_1 - \bar{x}_2|, \frac{s_1^2}{s_2^2}, \frac{IQR_1}{IQR_2}$, anything) and find its probability from the pooled sample (corresponding to the null hypothesis).



Wilcoxon-Mann-Whitney non-parametric test

This is a test developed before computers made the bootstrapping tests described above possible. It estimates whether ***the ranking*** of the values of two groups of data are significantly different, rather than the values themselves, so it can be applied to any type of data, without requiring a parametric distribution of the data.

There are two groups of size n_1 and n_2 and a total of $n = n_1 + n_2$
For the null hypothesis (that the two groups would have similar ranks) we pool the two groups and compute a total rank
$R = 1 + 2 + ... + n = n(n+1)/2$

We add up the rank of the elements of group 1 and group 2 when **pooled together** in the null hypothesis pool and get R_1, R_2, with $R_1 + R_2 = R$
It turns out that the statistic

$U = R - n(n+1)/2$ is Gaussian, with a mean $\mu_U = \frac{n_1 n_2}{2}$ and standard deviation $\sigma_U = \sqrt{\left[\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right]}$. So one computes the probability of getting
$U_1 = R_1 - n_1(n_1+1)/2$
within the null hypothesis distribution, checking on $Z = \frac{U_1 - \mu_U}{\sigma_U}$. If the probability is less than 5% (or 2.5% for a two tailed problem) we reject the null hypothesis.

Example 1: Assume the rankings for group 1 are 1,3,5,7,9, and for group 2 they are 2,4,6,8,10. What are their probabilities? Can we reject the null hypothesis? $\mu_U = 5*5/2 = 12.5; \sigma_U = \sqrt{5*5*11/12} = 4.79$; $Z_1 = 0.48$ $Z_2 = 0.52$. Obviously, values only 0.2 σ from the mean have high probability under the null hypothesis, which therefore cannot be rejected.

Example 2: Assume the rankings for group 1 are 1,2,3,4,5, and for group 2 they are 6,7,8,9,10. What are their probabilities? Can we reject the null hypothesis? $Z_1 = -2.61$ $Z_2 = +2.61$. Obviously, values 2.61 σ away from the mean have low probability under the null hypothesis, which therefore has to be rejected.

Hypothesis testing and Multiplicity Problem

Example: We make 20 independent tests at 5% level of significance, and two of them result positive, i.e., reject H_0. Should H_0 then be rejected, since 10% of the tests are positive? Actually not! Let's look at the probability of finding positive results in 20 independent tests if each one has only a 5% probability:

$$P(X=0)=\binom{20}{0}0.05^0 0.95^{19}=0.358$$

$$P(X=1)=\binom{20}{1}0.05^1 0.95^{19}=0.377$$

$$\Rightarrow P(X\geq 2)=1-.358-.377=0.265>0.05!$$

If the tests are not independent (e.g., grid points in the model) the multiplicity problem is even worse! One needs to do non-parametric tests for field significance (see section 5.4).

Exercise: Consider again the Guayaquil Table and test the hypotheses:
a) It's warmer during an El Niño
b) Pressure is lower during an El Niño
c) It rains more during an El Niño
Check level of significance p with which you can reject the null hypothesis.
Which of a), b), c) would be better to do with a nonparametric test?