MET3220C & MET6480
Computational Statistics

Hypothesis Testing
Parametric tests
(Chapter 5.2 of Wilks' book)

Key Points:
1) t-test
2) Tests for differences in means
3) Serial Correlation

http://campus.fsu.edu/
bourassa@met.fsu.edu

The Florida State University

Student's t-test

- The test is called Student's t-test.
  - The creator of this test published under the pseudonym Student.
  - His employer did not permit him to use his own name.
    - W. S. Gossett [1876-1937]
    - His real job was as a brew master.
- This test does not assume that the data distribution is Gaussian.
  - In practice, it is approximately Gaussian.
  - Gaussian distributions are common, as would be expected based on the Central Limit Theorem.
- The one-sample t-test is one of the most common statistical tests.
  - It examines differences between the mean of a sample population, and a previous specified mean (μ_0).
  - If there are enough data points the null distribution for a mean will be Gaussian (via the Central Limit Theorem).

Student's t-test

- The test statistic is

$$t = \frac{\bar{x} - \mu_0}{\left[Var(\bar{x})\right]^{1/2}}$$

  - Where Var indicates the variance.
- Note that the equation is similar to a *z* value, except that the standard deviation applies to the uncertainty in the mean, rather than the spread of the data, and the mean is not the mean of the sample.
- The other key consideration for this application is the degrees of freedom (ν).
  - For large degrees of freedom (a large number of independent observations), the Gaussian approximation is good.
  - $\nu = n - 1$, where n is the number of independent observations.
  - The variance of the mean of *x* is determined as σ_x^2/n, where σ_x is the variance of *x*.
- When the Gaussian approximation is good, Gaussian probabilities can be used to estimate the likelihood of a *t* value occurring, similar to using a *z* value. Otherwise tables are given in many stats books.

Tests for Differences in Means

- Another common application is examination of the difference between two independent means.
  - Example: differences in mean equatorial sea surface temperatures during different ENSO phases.
  - Example: differences in the number of mean land falling hurricanes during different phases of the Atlantic Multi-decadal Oscillation.
- The null hypothesis is often that the differences in the mean are plausible random differences. The alternative hypothesis is that the differences in the mean is larger than can be expected from random differences due to different samples.
  - This is a two tailed test.
  - The problem could also be setup as a one tailed test by specifying the sign of the difference in means.

Setting up the test

- The test statistic is

$$z = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - E\left[\overline{x}_1 - \overline{x}_2\right]}{\left(Var\left[\overline{x}_1 - \overline{x}_2\right]\right)^{1/2}}$$

- Where $E\left[\overline{x}_1 - \overline{x}_2\right] = E\left(\overline{x}_1\right) - E\left(\overline{x}_2\right) = \mu_1 - \mu_2$ is the expected difference in means. It is usually equal to zero for the null hypothesis, but could be equal to any finite value.
- The variance of the difference in the means could be calculated with the error propagation equation. It is

$$Var\left[\overline{x}_1 - \overline{x}_2\right] = Var\left[\overline{x}_2\right] + Var\left[\overline{x}_2\right]$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- The probability is then evaluated as for a z-value.

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}$$

Hurricane Example

Number of Atlantic tropical storms and hurricanes per year for 35 years.

1970	10	1983	4	1995	19
1971	13	1984	13	1996	13
1972	5	1985	11	1997	8
1973	7	1986	6	1998	14
1974	8	1987	7	1999	12
1975	8	1988	12	2000	14
1976	9	1989	11	2001	15
1977	6	1990	14	2002	12
1978	12	1991	8	2003	16
1979	8	1992	7	2004	14
1980	11	1993	8		
1981	11	1994	7		
1982	5				

- It has been claimed that since 1995 we have been in an active phase for tropical cyclones.
- The period of reduced activity is less clear, but if it is due to basin wide variability in the Atlantic Ocean, I claim it is from 1982 to 1997.
  - We will use 1982 to 1994 in this comparison.
- We are comparing two independent means.
- $\mathbf{H_0}$: difference in means is small.
- $\mathbf{H_A}$: difference in means is statistically significant.
- **Null distribution**: Gaussian

Hurricane Example

Number of Atlantic tropical storms and hurricanes per year for 25 years.

1982	5
1983	4
1984	13
1985	11
1986	6
1987	7
1988	12
1989	11
1990	14
1991	8
1992	7
1993	8
1994	7
1995	19
1996	13
1997	8
1998	14
1999	12
2000	14
2001	15
2002	12
2003	16
2004	14

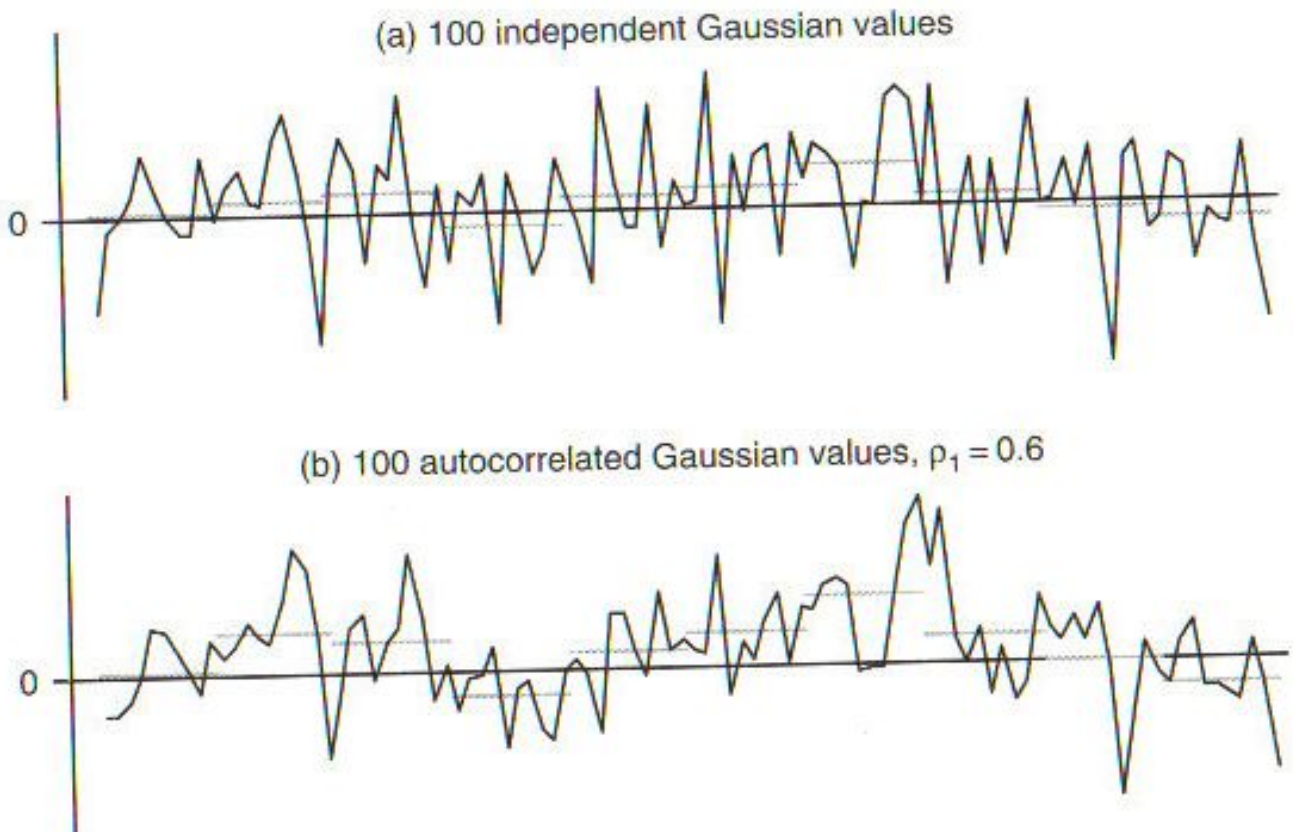
- Our null distribution, in more detail is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - E[\bar{x}_1 - \bar{x}_2]}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}} = \frac{\bar{x}_1 - \bar{x}_2}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}$$

- It remains to chose a confidence limit, determine the spread or uncertainty (standard deviation) in the test statistic, and calculate the z value.
- Let the 1982 to 1994 values be *set 1*, and the recent values be *set 2*.
- $n_1 = 13$, $\bar{x}_1 = 8.69$, and $s_1 = 3.04$
- $n_2 = 10$, $\bar{x}_2 = 13.7$, and $s_2 = 2.72$
- Results in an uncertainty of 1.20 tropical storms
- And z = –4.1
- Interpretation: null hypothesis is rejected with great confidence (<0.001% odds of false reject.).

Word of Caution
Interpretation Can Be More Complicated

- If the value at one time step is closely related to the value at adjacent time steps, then changes in the means as shown in the previous example are expected!



Differences of Means for Serially Dependent Data

- A great deal of meteorological data has some dependence on previous values.
  - Examples:
    - Hourly data showing a diurnal cycle,
    - Daily data showing a synoptic cycle,
    - Monthly averages showing an annual cycle.
- The problem of serial dependence can be avoided by sub-sampling data in a manner that puts sufficient time between the observations in the sub-sample.
  - Example: Taking every 10^{th} daily value over several years.
- Unfortunately, there are many applications where the time series is insufficiently long to avoid problems with serial dependence.
  - For example, a correlation between daily temperatures and the same time series lagged by one day (ρ_1) is around 0.6.

Some Characteristics of Serially Dependent Data

- How do statistics of serially dependent data differ from serially independent data?
  - A larger variance,
  - Adjacent (or nearby) data points tend to be more similar,
  - Serially dependent data are usually smoother,
  - Differences from the mean (over a short time period) are less likely to cancel, resulting in local averages that are farther from the true mean. Hence variances are likely to be larger than for serially independent data.

Working with Serially Dependent Data

- One way to deal with this problem is to determine an effective sample size (*n'*, based on independent observations) that is smaller than the actual sample size (*n*).

$$n' \approx n \frac{1-\rho_1}{1+\rho_1}$$

- If $\rho_1 = 0.6$, then $n'/n = (1–0.6)/(1+0.6) = 0.4 / 1.6 = 0.25$
- Similarly, the variance (or uncertainty squared) can be modified.

$$Var[\bar{x}] \approx \frac{s^2}{n'} = \frac{s^2}{n} \frac{1+\rho_1}{1-\rho_1}$$

- Inverting the ration can be thought of as the time (in units of the sampling interval) between independent observations.
- These equations apply only to data with only a lag 1 correlation. More complicated functions exist for more complicated relationships.

Example

- Consider a time series of daily maximum temperatures for January 1987 from two cities that are close together.
  - See table A.1 in Wilk's Appendix A for the values.
  - Answer the question 'are the means significantly different?'
- There are several ways to go about this.
  - Compute the means of each series, and determine if the differences is statistically different from zero.
  - Compute the differences and determine if the mean of the differences is statistically different from zero.
  - What are the pros and cons of each approach?
- Statistics that might be useful for determining which approach to use:
  - Lag 1 correlation is 0.52 for one city, and 0.61 for the other.
  - Lag 1 correlation is 0.076 for the differences.
  - Standard deviations are 7.71, 7.86, and 2.28°F.
  - Difference in the means (and mean of the differences) is –1.9°F.

Example Continued

- If we look at the difference in the means of each city, then we need to determine the number of independent data points for each city.
  - For the first city this is 31 (1 – 0.52) / (1 + 0.52) = 9.8 days.
  - For the 2nd city this is 31 (1 – 0.61) / (1 + 0.61) = 7.5 days.
- If we work with the mean of the paired differences, we use the same approach to determine the number of independent data points.
  - 31 (1 – 0.076) / (1 + 0.076) = 26.6
  - Substantially better than working with the individual cities.
- Recall that the difference in the means is the same in both approaches. The consideration that changes is the uncertainty, which is a function of the standard deviation (assuming a Gaussian distribution) and the number of independent data points.
- Recall that the standard deviation in the differences is approximately one third the standard deviations for the non-differenced values.

Example Concluded
The Test Statistics

- Consider approach using the mean of the temperature differences.
  - The difference in the means is –1.9°F
  - The uncertainty is $(2.28^2 / 26.6)^{1/2} = 0.442$°F
  - The z value is –4.29
- Consider the approach using the difference of the means.
  - The difference in the means is –1.9°F
  - The uncertainty is $(7.71^2 / 9.8 + 7.86^2 / 7.5)^{1/2} = 3.78$°F
  - The z value is –0.502
- Why is the approach based on the ‘mean of the differences’ so much better than the other approach?

  1) More independent points results in smaller uncertainty in the mean.

  2) The cities are closely located, so there is a high correlation between the temperatures. The variability associated with this correlation is removed from the differences, resulting in less uncertainty.