Review - 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 2^{nd} 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.

MET3220C & MET6480
Computational Statistics

Hypothesis Testing
Parametric tests:
Goodness of Fit

(Chapter 5.2.3 of Wilk's book)

Key Points:
1) χ^2 Test
2) K-S Test
3) Filliben Q-Q Correlation test

Goodness of Fit Tests

- In earlier lectures and assignments we discussed some simple tests.
  - Compare the range of the data to the range of the parametric distribution.
    - Are negative values found?
      - Can they be explained by random noise?
    - Are there other limiting values?
  - Compare a histogram of the data to the theoretical parametric distribution.
    - Use the data to determine the fitting parameters.

Distribution	$\mu = E(X)$	$\sigma^2 = Var[X]$	
Binomial	N p	N p (1 – p)	
Geometric	1/p	$(1 – p) / p^2$	
Negative Binomial	k (1 – p) / p	$K (1 – p) / p^2$	
Poisson	μ	μ	

Objective Measures of Goodness of Fit

- Goodness of Fit tests are unusual because the goal is often to support the null hypothesis.
  - Null Hypothesis: The data are consistent with the hypothesized distribution.
- Example of why we don't base our tests solely on the range of the data.
  - Scientists working with satellite observations of radar backscatter (the fraction of the radar signal that returns to the satellite), were greatly disturbed to find negative values. These values did not match modeled backscatter, which only allowed positive values.
  - Some people were ready to totally reject the basis for these models, and come up with a new distribution.
  - In reality, the negative values were consistent with very low signals and relatively large random error.
- It is better to focus on the distribution than subtle differences in acceptable bounds.

χ^2 Test

- The χ^2 test is a relatively common and relatively simple test for goodness of fit. It compares values in an observed histogram to values from a theoretical distribution.
- This test involve partitioning the data into bins.
  - Examples:
    - Probability of wind speeds in 0.5 m/s bins.
    - Probability of an annual number of landfalling tropical storms.
- The χ^2 test is much more natural for the second example, because the values are discrete, and easily binned. Rounding can be an issue when the technique is applied to continuous distributions.
- Continuous data should be integrated over each bin.

χ^2 Test Statistic

$$\chi^2 = \sum_{\text{bins}} \frac{(\#\text{observed} - \#\text{expected})^2}{\#\text{expected}}$$

$$\chi^2 = \sum_{\text{bins}} \frac{(\#\text{observed} - n\,\text{bin\_width}\Pr\{\text{data in bin}\})^2}{n\,\text{bin\_width}\Pr\{\text{data in bin}\}}$$

- If the model is a good fit to the data, then the χ^2 value will be 'small.'
- If the data is a poor fit, the χ^2 value will be much larger.
- The bins should span the entire range of the union of observations and theoretical values.
- It must be complete. We can't ignore the observations that don't fit the model!
- The number of degrees of freedom (ν) is:
  ν = number of bins – number of fitting parameters – 1
- The test is always one sided.
- Likelihoods are given in Wilk's Table B.3.

χ^2 Table: Confidence Limits for Rejection of Null Hypothesis

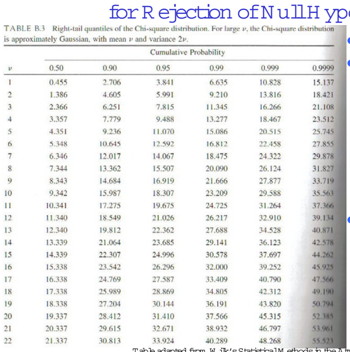


TABLE B.3 Right-tail quantiles of the Chi-square distribution. For large ν, the Chi-square distribution is approximately Gaussian, with mean ν and variance 2ν.

	Cumulative Probability					
ν	0.50	0.90	0.95	0.99	0.999	0.9999
1	0.455	2.706	3.841	6.635	10.828	15.137
2	1.386	4.605	5.991	9.210	13.816	18.421
3	2.366	6.251	7.815	11.345	16.266	21.108
4	3.357	7.779	9.488	13.277	18.467	23.512
5	4.351	9.236	11.070	15.086	20.515	25.745
6	5.348	10.645	12.592	16.812	22.458	27.855
7	6.346	12.017	14.067	18.475	24.322	29.878
8	7.344	13.362	15.507	20.090	26.124	31.827
9	8.343	14.684	16.919	21.666	27.877	33.719
10	9.342	15.987	18.307	23.209	29.588	35.563
11	10.341	17.275	19.675	24.725	31.264	37.366
12	11.340	18.549	21.026	26.217	32.910	39.134
13	12.340	19.812	22.362	27.688	34.528	40.871
14	13.339	21.064	23.685	29.141	36.123	42.578
15	14.339	22.307	24.996	30.578	37.697	44.262
16	15.338	23.542	26.296	32.000	39.252	45.925
17	16.338	24.769	27.587	33.409	40.790	47.566
18	17.338	25.989	28.869	34.805	42.312	49.190
19	18.338	27.204	30.144	36.191	43.820	50.794
20	19.337	28.412	31.410	37.566	45.315	52.385
21	20.337	29.615	32.671	38.932	46.797	53.961
22	21.337	30.813	33.924	40.289	48.268	55.523

Table adapted from Wilk's Statistical Methods in the Atmospheric Sciences

- Wilk's Table B.3
- The table gives the minimum χ^2 values required to reject the null hypothesis, as a function of the confidence limit and the number of degrees of freedom.
- For large values of ν, the distribution is approximately Gaussian, with a mean of ν and a standard deviations of 2ν.

Example: Comparing Gaussian and Gamma Distributions to Data

- Consider 50 years of January precipitation in Ithica (Wilk's example 5.3).
  - Compare the fits based on Gamma and Gaussian distributions.
- Determine the fitting parameters for each distribution:
  - α = 3.76 and β = 0.52 for the Gamma distribution, and
  - Mean of 1.96" and standard deviation of 1.12"
- Integrate PDFs over each bin, and multiply by 50 to get number of observations.

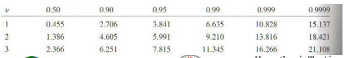
bins	<1"	1–1.5"	1.5 – 2"	2 – 2.5"	2.5 – 3"	≥3"
Observed number	5	16	10	7	7	5
Gamma Distrib.:						
Probability	0.161	0.215	0.210	0.161	0.108	0.145
Expected	8.05	10.75	10.50	8.05	5.4	7.25
Gaussian Distrib.:						
Probability	0.195	0.146	0.173	0.173	0.132	0.176
Expected	9.75	7.30	8.65	8.90	6.60	8.80

Example: Comparing Gaussian and Gamma Distributions to Data

- Calculate the χ^2 values for each distribution.

$$\chi^2 = \sum_{\text{bins}} \frac{(\#\text{observed} - \#\text{expected})^2}{\#\text{expected}}$$

- χ^2 values are:
  - 5.05 for the Gamma distribution
  - 14.96 for the Gaussian distribution
- The number of degrees of freedom:
  - ν = number of bins – number of fitting parameters – 1
  - ν = 6 – 2 – 1 = 3
- The null hypothesis would be
  - Not rejected at the 90% confidence limit for the Gamma distribution,
  - Rejected at the 99% limit, but not the 99.9% limit for the Gaussian distribution.



ν	0.50	0.90	0.95	0.99	0.999	0.9999
1	0.455	2.706	3.841	6.635	10.828	15.137
2	1.386	4.605	5.991	9.210	13.816	18.421
3	2.366	6.251	7.815	11.345	16.266	21.108

Word of Caution

- The χ^2 test does not consider the consequences of uncertainty (random errors) in the observations. If these errors are large compared to the bin width, the χ^2 test could be very misleading!

Kolmogorov-Smirnov and Lilliefores Tests

- The K-S test is another commonly applied test for goodness of fit.
  - Recall that the χ^2 test compared the observed and modeled PDFs.
  - In contrast, the K-S test examines the observed and modeled CDFs.
- Null hypothesis: the modeled data is a statistically good fit to the observed data.
  - If our test statistic is too large, then the null hypothesis is rejected.
- Note that the K-S test is usually a more sensitive test than the χ^2 test.
  - Particularly so for continuous distributions.
- Note that the K-S test is not applicable when the fitting parameters are determined from the observations.
  - Since fitting parameter are often determined in this fashion, this consideration is a very serious constraint!
- The modified K-S test, or Lilliefores test (Lilliefores 1967), can be (correctly) applied when the fitting parameters are determined from the observations.

Kolmogorov-Smirnov and Lilliefores Tests

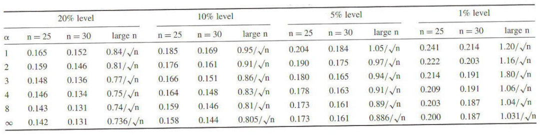
- The test statistic is the absolute value of the largest difference between the observed and modeled CDF.
  - The differences are calculated only for the CDF values corresponding to each observation.
  - Note that the data does not have to be binned
- The test statistic can be written as $D = \max\left|CDF_{obs}(x_i) - CDF_{model}(x_i)\right|$,
  - Where $CDF_{obs}(x_i) = i/n$
  - Where x_i is the i^{th} smallest value. Think of the series x as sorted from smallest to largest values.
- The test statistics (D) is compared to a critical (C) value that is a function of the confidence limit and the number of observations (n).
  - If $D \geq C$ then the null hypothesis is rejected.

$$C = \frac{K}{\sqrt{n} + 0.12 + 0.11\sqrt{n}}$$,For the K-S test only.

  - Where $K = 1.224, 1.358$, and 1.628 for $\alpha = 0.10, 0.05$, and 0.01.

Lilliefores Tests Critical Values

- The critical values for the Lilliefores test are dependent on the theoretical distribution.
- Critical values have been determined (Crutcher 1975) for Gamma distributions
  - Recall that α is the shape parameter.



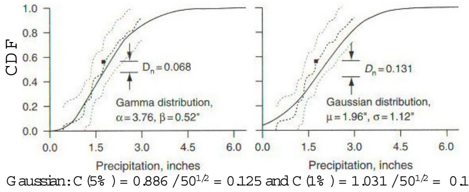
	20% level			10% level			5% level			1% level		
α	n = 25	n = 30	large n	n = 25	n = 30	large n	n = 25	n = 30	large n	n = 25	n = 30	large n
1	0.165	0.152	$0.84/\sqrt{n}$	0.185	0.169	$0.95/\sqrt{n}$	0.204	0.184	$1.05/\sqrt{n}$	0.241	0.214	$1.20/\sqrt{n}$
2	0.159	0.146	$0.81/\sqrt{n}$	0.176	0.161	$0.91/\sqrt{n}$	0.190	0.175	$0.97/\sqrt{n}$	0.222	0.203	$1.16/\sqrt{n}$
3	0.148	0.136	$0.77/\sqrt{n}$	0.166	0.151	$0.86/\sqrt{n}$	0.180	0.165	$0.94/\sqrt{n}$	0.214	0.191	$1.80/\sqrt{n}$
4	0.146	0.134	$0.75/\sqrt{n}$	0.164	0.148	$0.83/\sqrt{n}$	0.178	0.163	$0.91/\sqrt{n}$	0.209	0.191	$1.06/\sqrt{n}$
8	0.143	0.131	$0.74/\sqrt{n}$	0.159	0.146	$0.81/\sqrt{n}$	0.173	0.161	$0.89/\sqrt{n}$	0.203	0.187	$1.04/\sqrt{n}$
∞	0.142	0.131	$0.736/\sqrt{n}$	0.158	0.144	$0.805/\sqrt{n}$	0.173	0.161	$0.886/\sqrt{n}$	0.200	0.187	$1.031/\sqrt{n}$

- For Gaussian distributions use the $\alpha = \infty$ row.

Table adapted from Wilk's Statistical Methods in the Atmospheric Sciences

Lilliefores Test Example

- The Lilliefores test is applied to two fits to precipitation observations
  - Observations are from Ithica in January
  - Theoretical distributions are Gamma and Gaussian.
  - n = 50 (large n column on previous table)



Gaussian: C(5%) = $0.886/50^{1/2} = 0.125$ and C(1%) = $1.031/50^{1/2} = 0.146$

Gamma: C(20%) = $0.75 / 50^{1/2} = 0.106$

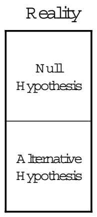
Why is a 20% chance of false rejection better than a 5% chance?

Figure adapted from Wilk's Statistical Methods in the Atmospheric Sciences

Review of Hypotheses

- The null hypothesis (H_0).
  - The null hypothesis is part of a logical structure which is used to examine the test statistic.
  - The null hypothesis is often designed as the compliment to what we would like to test for.
    - Example: student A is not statistically taller than student B.
    - Example: Any rate of temperature change is either negative or 'positive and statistically indistinguishable from zero'.
- The alternative hypothesis (H_A).
  - This hypothesis is the compliment of the null hypothesis.
    - Example: the null hypothesis in not true.
  - A more complicated hypothesis is possible.
  - Hint: think about whether it is easier to clearly state null hypothesis or an alternative hypothesis, then define the other hypothesis as the compliment of the one that is more easily defined.

Graphical Example of Hypotheses



Reality

Null Hypothesis

Alternative Hypothesis

- The null hypothesis together with the alternative hypothesis describe all outcomes related to the question.
- The boxes to the left describe two possible states:
  - The null hypothesis is correct, or
  - The alternative hypothesis is correct.
  - There is no alternative outcome, and both hypotheses cannot be correct.
- Data analysis (AKA statistics) can be used to try to determine which of the hypotheses is true.

Graphical Example of Perception or example of statistical inference

Statistical Inference

Null Hypothesis is accepted
Alternative Hypothesis is accepted

- Statistical testing can be used to either accept or reject the null hypothesis.
  - Reject the null hypothesis is the same as accept the alternative hypothesis.
- The boxes to the left describe two possible outcome of the statistical test:
  - The null hypothesis is accepted, or
  - The alternative hypothesis is accepted.
  - There is no alternative outcome, and both hypotheses cannot be correct.
- So where does statistical confidence (e.g., the chance of a false rejection of the null hypothesis) come into play?
  - You need to combine the reality and statistical inference cases.

Putting it Together

Reality About Null Hypothesis	
True	
False	

- In reality, there are two possibilities about the null hypothesis:
  - It is true, or
  - It is false.

Putting it Together

Statistical Inference About Null Hypothesis

Reality About Null Hypothesis	Accepted	Rejected
True		
False		

- In reality, there are two possibilities about the null hypothesis:
  - It is true, or
  - It is false.
- There are also two possibilities for our statistical inference (perception) about the null hypothesis:
  - It is true (accepted), or
  - It is false (rejected).

Putting it Together

Statistical Inference About Null Hypothesis

Reality About Null Hypothesis	Accepted	Rejected
True		
False		

- In reality, there are two possibilities about the null hypothesis:
  - It is true, or
  - It is false.
- There are also two possibilities for our statistical inference (perception) about the null hypothesis:
  - It is true (accepted), or
  - It is false (rejected).
- Combining reality with perception results in 4 (2x2) possible outcomes:
  - Correct acceptance,
  - Correct rejection,
  - False rejection (type I error), and
  - False acceptance (type II error).

Putting it Together

Statistical Inference About Null Hypothesis

Reality About Null Hypothesis	Accepted	Rejected
True	Correct Acceptance	False Rejection (α)
False	False Acceptance (β)	Correct Rejection

- In reality, there are two possibilities about the null hypothesis:
  - It is true, or
  - It is false.
- There are also two possibilities for our statistical inference (perception) about the null hypothesis:
  - It is true (accepted), or
  - It is false (rejected).
- Combining reality with perception results in 4 (2x2) possible outcomes:
  - Correct acceptance,
  - Correct rejection,
  - False rejection (type I error), and
  - False acceptance (type II error).

Thinking About It

Statistical Inference About Null Hypothesis

Reality About Null Hypothesis	Accepted	Rejected
True	Correct Acceptance	False Rejection (α)
False	False Acceptance (β)	Correct Rejection

- In this application, we typically take the Reality true or false to be either zero or one.

- It we have a 5% chance of a false rejection, is the chance of correct acceptance equal to 95% ?
- If so, then the odds of a correct acceptance plus the odds of a false rejection is equal to one.
- Clearly the answer is no, because that would mean there is NO CHANCE of a false acceptance or a correct rejection.
- The odds of
  - Correct acceptance, plus
  - Correct rejection, plus
  - False rejection, plus
  - False acceptance

  are equal to one.

Q-Q Test

- A very robust test for a Gaussian distribution is often needed.
- Comparisons based on the correlation of the observation-based quantiles and Gaussian quantiles are very robust (D'Agustino 1986).

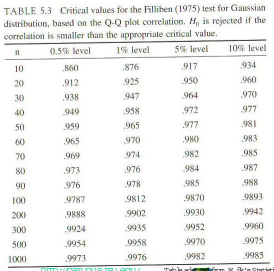


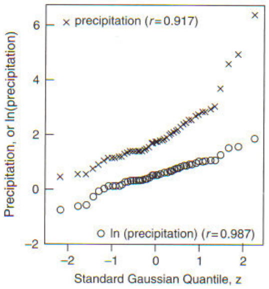
TABLE 5.3 Critical values for the Filliben (1975) test for Gaussian distribution, based on the Q-Q plot correlation. H_0 is rejected if the correlation is smaller than the appropriate critical value.

n	0.5% level	1% level	5% level	10% level
10	.860	.876	.917	.934
20	.912	.925	.950	.960
30	.938	.947	.964	.970
40	.949	.958	.972	.977
50	.959	.965	.977	.981
60	.965	.970	.980	.983
70	.969	.974	.982	.985
80	.973	.976	.984	.987
90	.976	.978	.985	.988
100	.9787	.9812	.9870	.9893
200	.9888	.9902	.9930	.9942
300	.9924	.9935	.9952	.9960
500	.9954	.9958	.9970	.9975
1000	.9973	.9976	.9982	.9985

- Filliben (1975) developed a simplified (and almost as effective) test.
  - The data value of the data are plotted on the dependent axis, and the Gaussian quantiles are plotted on the independent axis.
- The null hypothesis, that the data has a Gaussian distribution, is rejected if the correlation is less that the value in the table.

Table adapted from Wilk's Statistical Methods in the Atmospheric Sciences

Example of Filliben Q-Q Test



- Examine if the January precipitation data from Ithica can be modeled as Gaussian.
- Test Stat: Q-Q correlation
  H_0: Gaussian distribution
  H_A: Not Gaussian
  Confidence: 5% false rejection
  Null Distribution: from table, critical value of 0.977.
- Result: H_0 rejected
- Examine if the distribution is log-normal, with the same test, except that we are now examining log(precip).
- Result: H_0 accepted.