MONASH University

Information Technology

FIT5190 Introduction to IT Research Methods

Lecture 9
Quantitative Data Analysis
– Distributions and Hypothesis Testing

Slides prepared by
David Green, Frada Burstein, Jacques Steyn, Geoff Webb, Chung-Hsing Yeh

Learning objectives

- Understand
  - Probability distributions
  - Parametric versus non-parametric hypothesis testing
  - Type I and type II errors

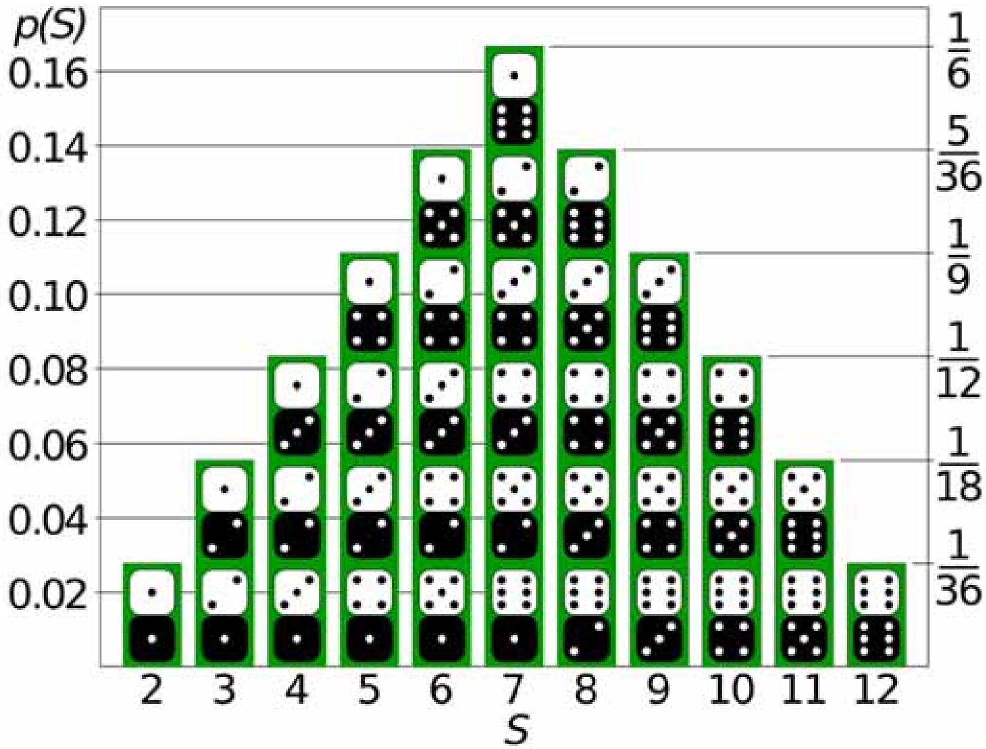
- Be able to
  - Apply a *t*-test on sample means
  - Apply a sign test on sample medians

Overview

- This lecture extends the discussion of hypothesis testing using elementary probability and statistics introduced in the last lecture to include topics such as
  - Distributions,
  - parametric and non-parametric tests,
  - hypothesis testing using the t-test and types of errors.

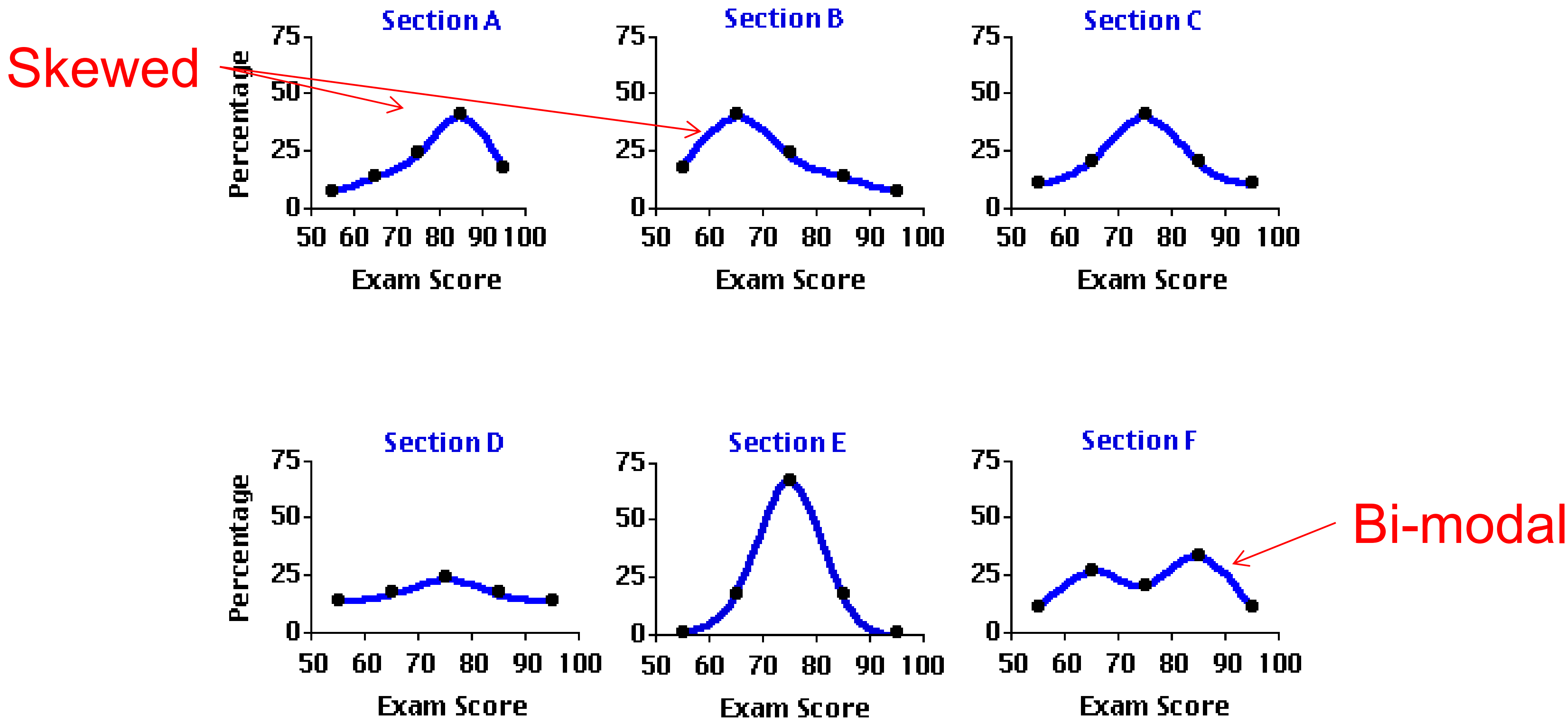
Probability distributions

- A probability distribution is a function that describes the probability of each value of a variable.
- Example: If we roll 2 dice, what is the probability distribution for the sum of their values?



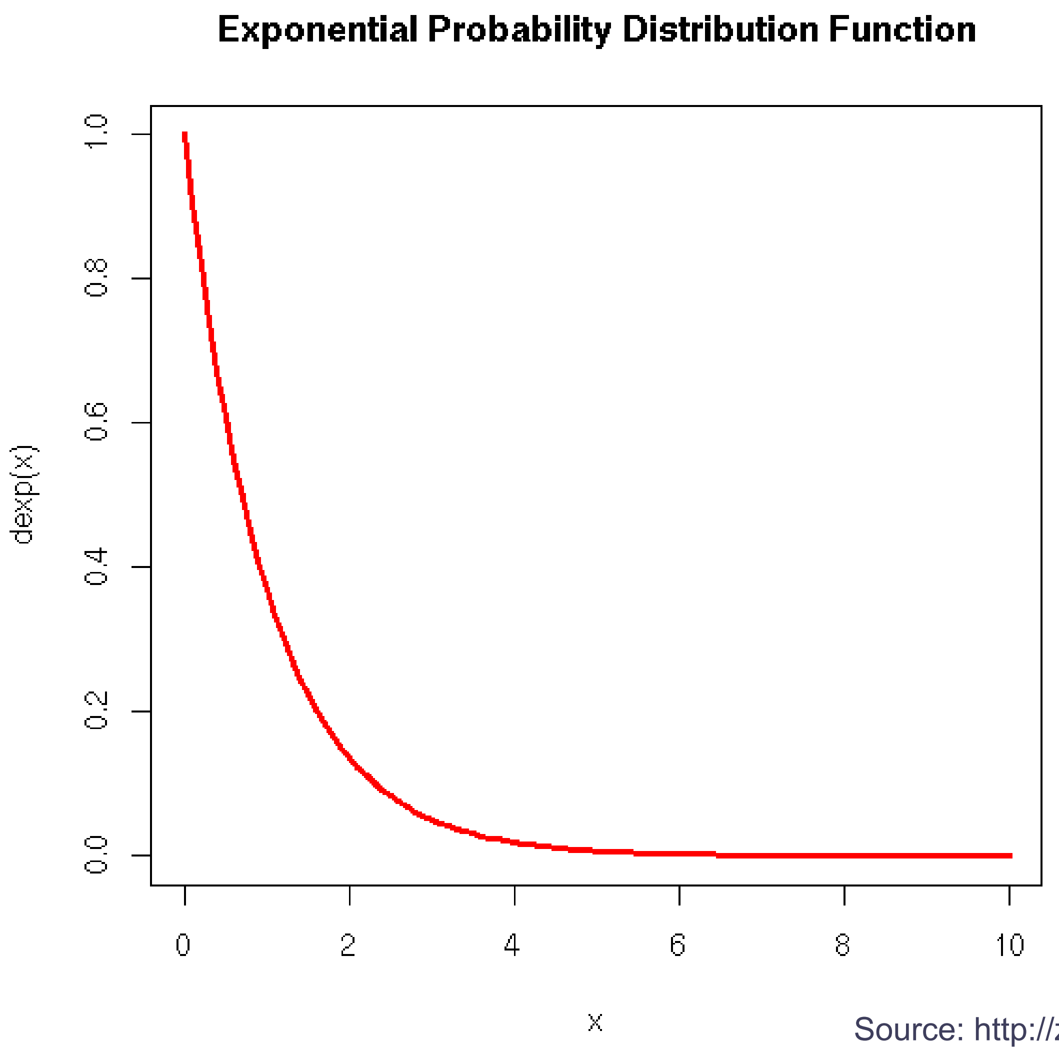
Source: http://en.wikipedia.org/wiki/File:Dice_Distribution_(bar).svg

More examples



Distributions of Exam Scores in Six Sections of a Statistics Course
Source: http://vassarstats.net/textbook/ch2pt1.html

Another example: exponential distribution

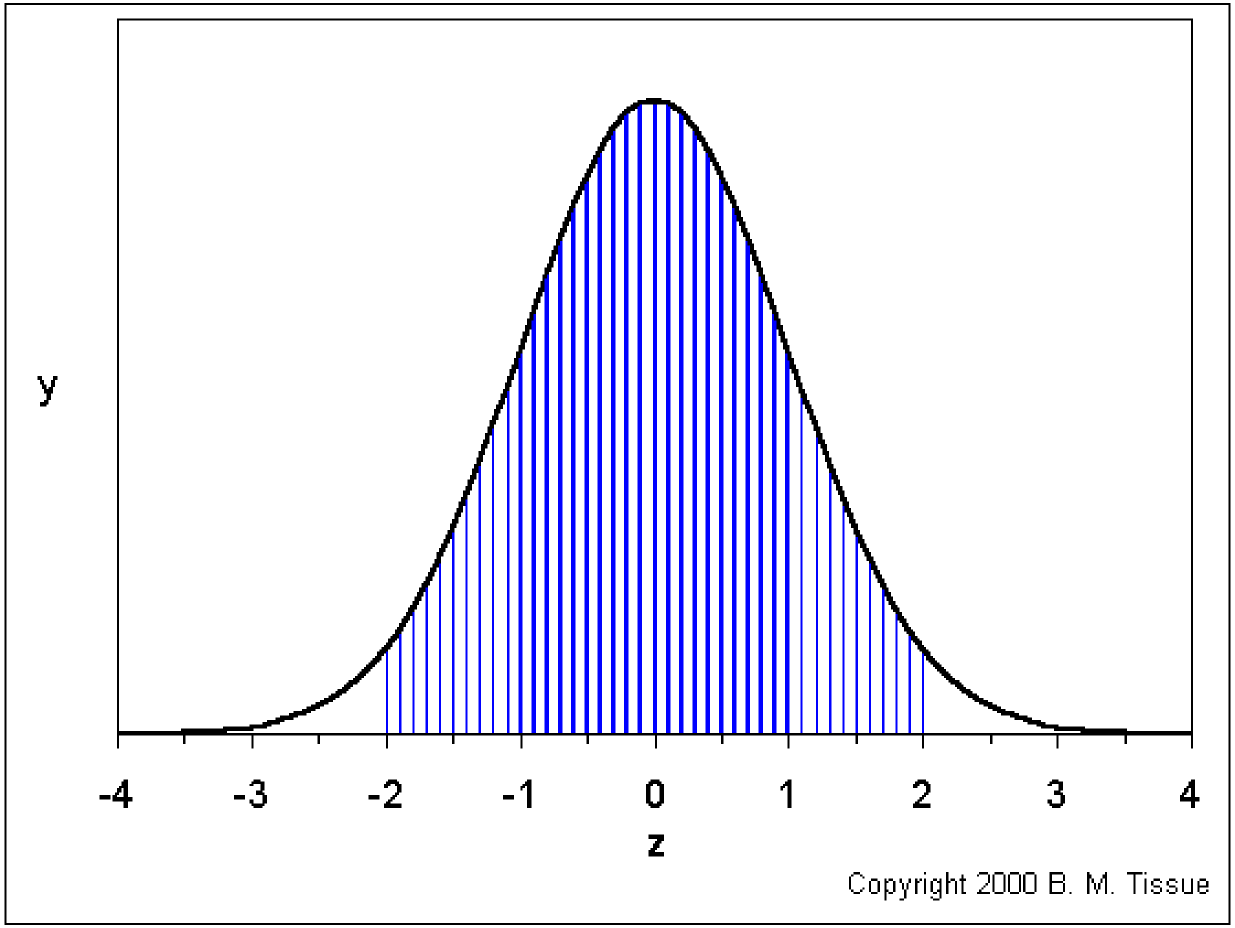


Source: http://zoonek2.free.fr/UNIX/48_R/07.html

Describes the time between events which occur continuously and independently at a constant average rate

Standardised distributions: Gaussian

- Also known as the **Normal Distribution**.
- Values cluster around a central value called the ***mean***.



Source: http://www.chemicool.com/definition/gaussian_distribution.html

Parameters of a distribution

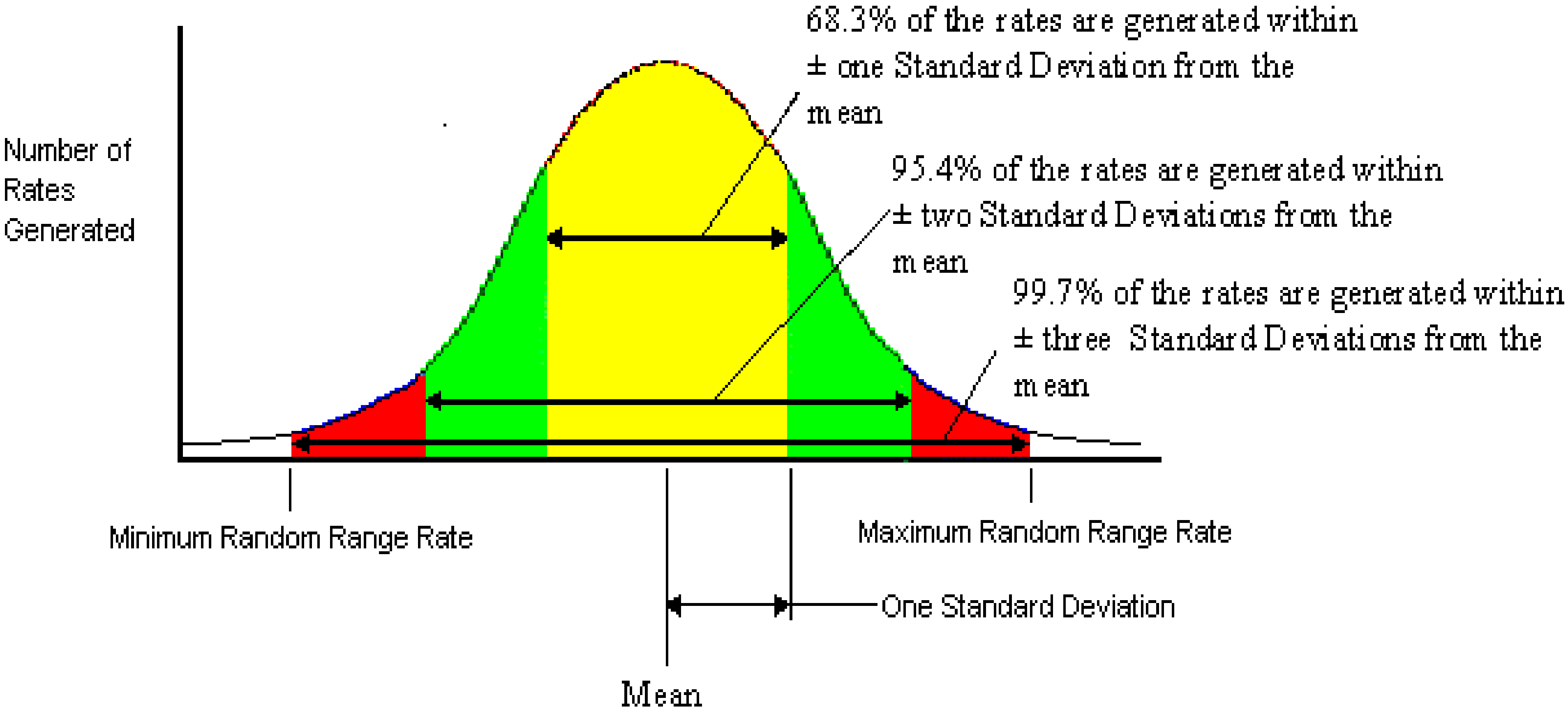
- Parameters of a distribution
  - Summarise the distribution shape, position, etc.

- Example:

  Parameters of the Normal (Gaussian) distribution:

  - Mean (average)
  - Variance (standard deviation squared)

Mean and standard deviations



Source: http://www.jlplanner.com/html/gauss.html

Note: variance (σ^2) is standard deviation (σ) squared

Mean, mode and median

- The *mean*, *mode* and *median* are three forms of *average.*
  - Measures of *central tendency*
- Mean
  - Most commonly used
  - Sum of values divided by number of values
- Mode
  - The most frequent value
- Median
  - The most central value

Mean, mode and median

- The *mean*, *mode* and *median* are three forms of *average.*
  - Measures of *central tendency*
  - Tell you about the *typical case*
- Mean
  - Most commonly used
  - Sum of values divided by number of values
- Mode
  - The most frequent value
- Median
  - The middle value

- Example:

  1, 2, 2, 2, 4, 4, 8, 8, 16
- Mean = 5.22
- Mode = 2
- Median = 4

Measures of spread

- Averages don’t tell you how often and how much the population differs from the average
- *Variance* and *standard deviation* tell you this
- Variance = mean squared deviation from the population mean
- Standard deviation = square root of variance

Central Limit Theorem (CLT)

- Under repeated sampling (with a sufficiently large number of observations generated randomly and independently), the sample means approximate a normal distribution.
  - Its mean equals the sample mean.
  - Its standard deviation is $\sigma/\sqrt{n}$, where n is the number of samples.

- The CLT means that we can base hypothesis tests on the normal distribution, **no matter what the underlying distribution may be.**

Parametric hypothesis testing

- Parametric hypothesis tests are often used to measure the quality of sample parameters or to test whether estimates on a given parameter are equal for two samples.
- Most common distribution parameters
  - Mean (average)
  - Variance (standard deviation squared)
- Test for the mean of a sample
- Test for means of two samples

Reminder: statistical hypothesis tests

- A statistic is a number calculated from a sample of observations
  - e.g. Average crop size
- Assuming the null hypothesis H_0
  - What is the probability of getting that number (or a more extreme value) purely by chance?
    - e.g. What is the probability of getting a bigger crop by chance?
- Significance
  - The probability that the result is not chance

Statistical hypothesis testing - summary

- The method
  - Use a set of observations
  - Set a "significance level" α (acceptance level) e.g. 0.01 or 0.05
  - Assuming the null hypothesis H_0 is true, calculate the probability p_0 of getting the observed result (p-value)
    - If $p_0 > \alpha$ retain the null hypothesis
    - If $p_0 < \alpha$ reject the null hypothesis

Example: plant fertilizer experiment #2

- Do we accept or reject an hypothesis?
  - H_1 : "Fertilizer makes a difference to crop size"
  - H_0 : "Fertilizer makes no difference to crop size"
- Convert the hypothesis into a statement about the *distribution* of observations
  - The average size of apples from fertilized trees will differ to those from unfertilized trees
  - H_1 : Average(X | fertilized) $\neq$ Average(X | unfertilized)
  - H_0 : Average(X | fertilized) = Average(X | unfertilized)

Populations and samples

- We use samples to estimate statistical parameters of populations

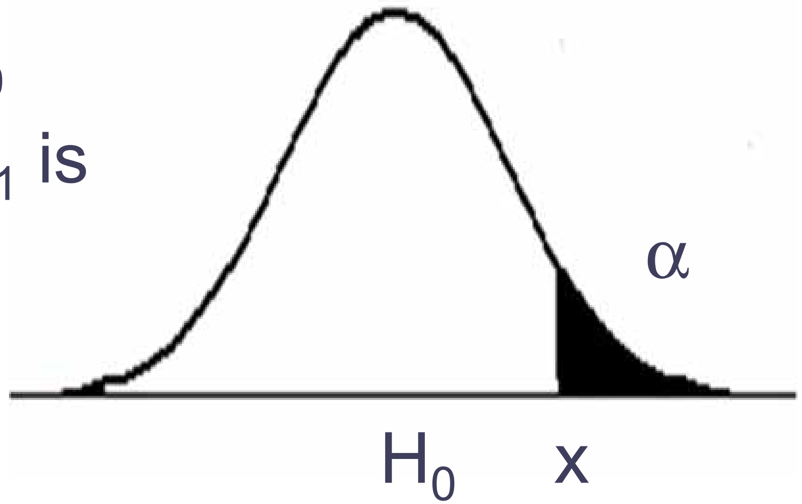
Population	
Size	*large*
Mean	μ
Variance	σ^2
Statistic	e.g. $Z \sim N(\mu,\sigma)$

Sample	
Size	N
Mean	$\bar{x}$
Variance	s^2
Statistic	t

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x}_1)^2$$

Significance tests

1. Set a significance level α (e.g. 0.05)
2. Compute a suitable statistic X from a sample
3. Compute $p(|x|>|X| \mid H_0)$ **
   – Probability of getting the result or a more extreme result if the null hypothesis H_0 is true
4. If $p(|x|>|X| \mid H_0) < \alpha$, then reject H_0
   – i.e. the alternative hypothesis H_1 is "significant" (acceptable)
   – α is the "rejection region"
5. It not, then reject H_1



** |x| means absolute value of x (ignores negation signs)
H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$

Hypotheses about the mean μ

One tailed test

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

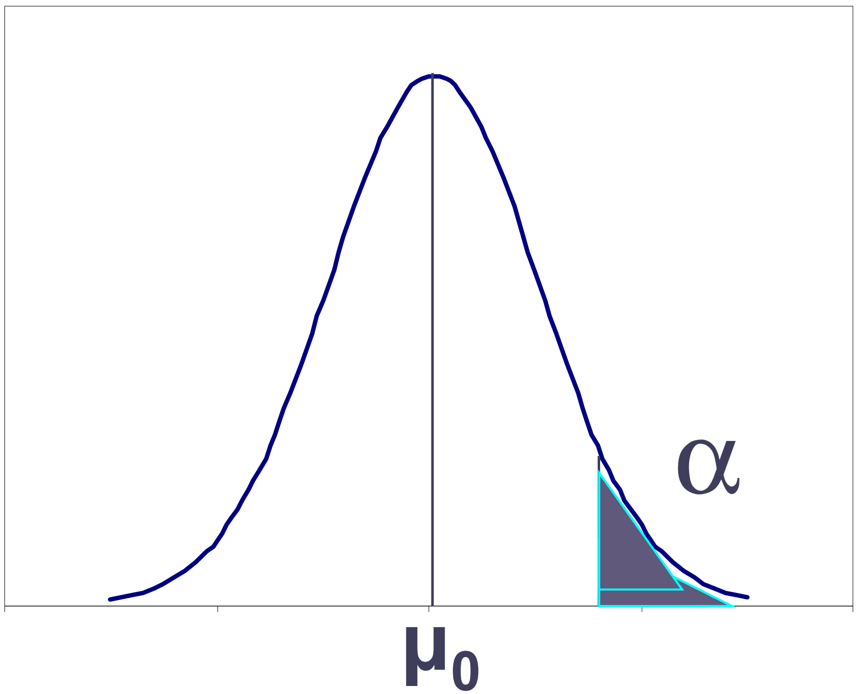
H_1: The mean of the sample is greater than the mean of the population

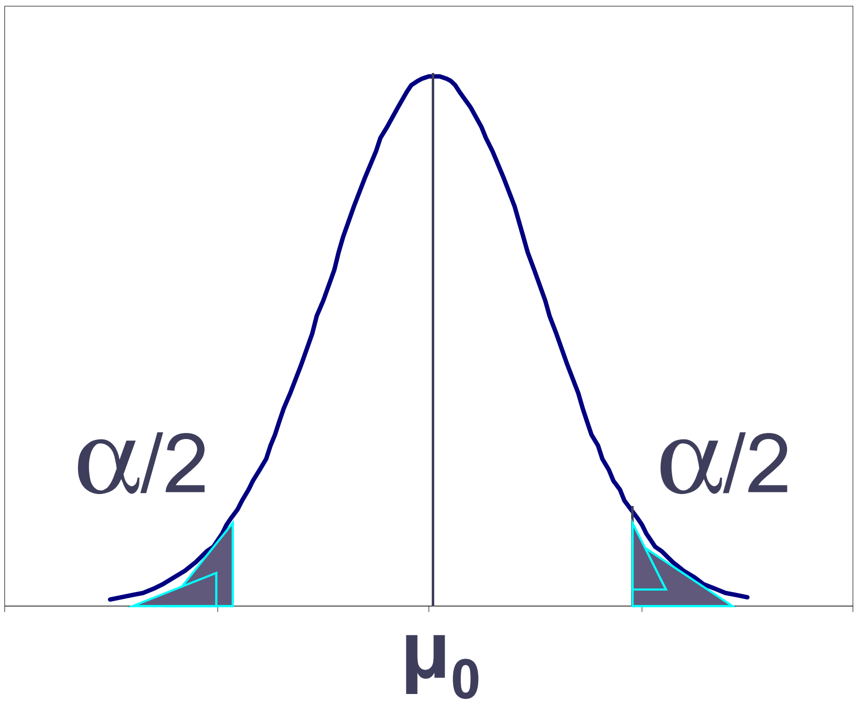
Two tailed test

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

H_1: The mean of the sample is different from the mean of the population





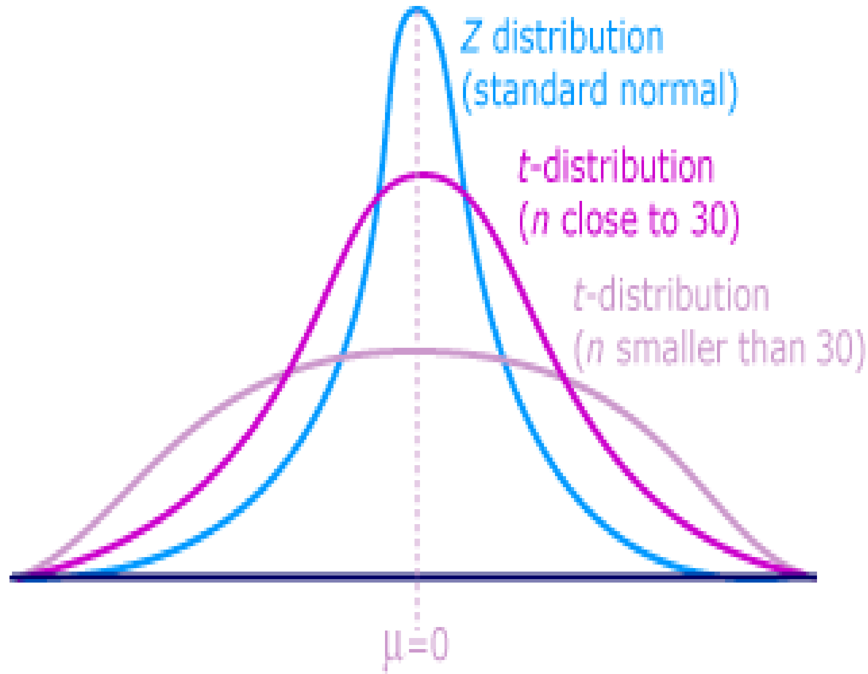
α = significance level

Student's *t*-distribution

- The *t*-distribution expresses the way in which sample means vary about the population mean.
- Use the *t*-test when the population variance is not known.
- The statistic *t*, given by

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

follows a *t*-distribution.



Fatter tails than normal distribution = higher probability of large values

t-test on the mean of a sample

- Null hypothesis
  - $\mu = \mu_0$ (assumes the population mean is known)
- Alternative hypothesis
  - One tailed test $\mu > \mu_0$ (or $\mu > \mu_0$)
  - Two tailed test $\mu \neq \mu_0$
- Test statistic

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \qquad df = n - 1$$

Practical notes re calculations

- Spreadsheets provide several functions for calculating *t* in different ways:

  TDIST returns value of *p* given *t* and *df*

  TINV returns value of *t* given *p* and *df*

  TTEST performs a *t*-test between two samples, returning the *p* value

- Be careful to check whether a test calculates values for a 1 or 2 sided test.

- Different authors calculate values in different ways (e.g. divide by *n-1* instead of *n*). So cite a reference for the version you use.

Example: *t*-test on the mean of a sample

- Improvements to an algorithm must take significantly less than 3000 seconds to solve a complex problem.
- A trial sample of 8 runs gave these results:

3005	**2925**	**2935**	**2965**	**2995**	**3005**	**2935**	**2905**

- Hypothesis test
  - H_0: $\mu = 3000$; H_1: $\mu < 3000$
  - Set $\alpha = 0.05$
  - $df = 8 - 1 = 7$
  - Critical t value required to reject H_0: $t(0.05, 7) = 1.895$
  - So we require $t < -1.89$ (mean must be $< \mu$, so negative)
  - $t = (\bar{x} - \mu_0)/s\sqrt{n} = (2958.75 - 3000)/(39.26 \times 2.83) = -0.37$
  - So retain null hypothesis; i.e. no improvement

t-test for difference between means of two samples

- Null hypothesis
  - $\mu_1 = \mu_2$ (i.e. no population mean is known)
- Alternative hypothesis
  - One tailed test $\mu_1 > \mu_2$ (or $\mu_1 < \mu_2$)
  - Two tailed test $\mu_1 \neq \mu_2$
- Test statistic for $(\mu_1 - \mu_2)$ $(\bar{X}_1 - \bar{X}_2)$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad df = n_1 + n_2 - 2$$

t-test for difference between sample means

- Since the population variance is unknown, we have to use the variance from the two samples i.e. $\sigma_{\bar{X}_1 - \bar{X}_2}$
- In general, the variances and sample sizes are different, so we use the formula

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

- If the samples are the same size, this simplifies to

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sum_i (S_1^2 + S_2^2)/(n-1)}}$$

Plant fertiliser experiment

- **Categorical approach**
  - Classify each plant as “large” or “small”
  - Use **chi-square** to test hypothesis about size

- **Scalar approach**
  - Measure size of each plant
  - Use ***t*-test** to test hypothesis about size

Example: plant fertilizer experiment #3

Fertilized	*Unfertilized*
32	35
37	31
35	29
28	25
41	34
44	40
35	27
31	32
34	31

H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$

$\text{mean}_1 = 35.22$, $s_1 = 4.94$

$\text{mean}_2 = 31.56$, $s_2 = 4.48$

Pooled $s^2 = (195.56+160.33)/(9+9-2) = 22.24$

so $s = 4.72$

$t = (35.22-31.56)/4.72*\text{sqrt}(2/9) = 1.64$

Set $\alpha = 0.05$, and $df = 9+9-2 = 16$

t required $= t(0.05,16) = 1.746$

Conclusion:

1. Retain null hypothesis
2. Fertilization does not increase size

Parametric vs. non-parametric hypothesis testing

Parametric statistics

- Parametric statistics are based on an assumption that the population from which the sample(s) are drawn conform to a specific distribution.
- e.g. the *t*-test assumes that the population(s) conform to normal distributions.
- e.g. if assume population *A* is a normal distribution with mean μ and standard deviation σ and another case is greater than $\mu + 3\sigma$ then there is less than 0.3% chance that the new case belongs to *A.*

Strengths of parametric statistics

- Typically easy to calculate
  - *t*-test only requires simple calculations using the mean and variance.
- Powerful
  - Requires relatively little data to reject false null hypothesis.

Limitations of parametric statistics

- More likely to falsely reject the null hypothesis if the assumptions made are incorrect.

Non-parametric statistics

- Non-parametric statistics make no assumptions about the form of distribution(s) from which the data are drawn.
- e.g. the chi-square test.

Strengths of non-parametric statistics

- As non-parametric methods make **fewer assumptions**, their applicability is much **wider** than the corresponding parametric methods.
- They can be applied safely in situations where **less is known** about the application in question.
- Non-parametric methods may be **easier to use**.
- Non-parametric methods leave **less room for improper** use and misunderstanding.

Limitations of non-parametric statistics

- Non parametric statistics are **“less powerful”** than parametric statistics.
  - They require **larger samples** to yield the same level of significance.
  - They are **less likely** to reject the null hypothesis.

Matching parametric and non-parametric tests

- Most parametric tests have a counterpart non-parametric test.
- e.g. *t*-test and Mann–Whitney U test
  - also known as Mann–Whitney–Wilcoxon (MWW) or Wilcoxon rank-sum test
- *t*-test tests whether **means** of two populations differ.
- Mann–Whitney U test tests whether the **medians** differ.

Test assumptions

- Every statistical test is based on a number of assumptions about the data.
  - e.g. the form of distribution
- It is important to check that the assumptions of a test hold for your data.
- One assumption made by most tests is that all data points are independent of each other.
  - True for coin tosses
  - Not true for most time series

Matched-pairs tests

- Most tests assume that the values in one treatment are independent of those in the other(s).
- Matched pairs tests assume that they are related.
  - Pre and post tests (i.e. Before and After tests)
  - Performance of two different populations on each of many tasks

Before	After
2	3
3	2
2	4

Matched-pairs tests

Before	After
2	3
3	2
2	4

- The matched-pairs *t*-test
  - Two tailed null hypothesis
    - The means are identical.
  - One tailed null hypothesis
    - One specified mean is greater than the other.
  - Assumes that the differences are normally distributed.
    - Unlikely to be true unless the data are ratio scale.

Matched-pairs tests

Before	After
2	3
3	2
2	4

- The sign test (non-parametric)
  - One tailed null hypothesis
    - Medians are identical.
  - Two tailed null hypothesis
    - The median of one specified population is greater than that of the other.
  - Assumptions
    - The values are on the same scale hence comparable
    - The values for one pair are independent of the values of another.

Sign test method

- If the medians are identical then p(X > Y) = 0.5.
- Count the number of pairs for which each treatment has the higher value.
  - Ignore those cases where the values are the same.
- Under the null hypothesis, this will follow the binomial distribution, which can be used to calculate the probability of obtaining the observed difference by chance.

Before	After
2	3
3	2
2	4

Example

Company image before and after viewing web site

Image before	Image after	Image before	Image after
2	3	6	5
3	2	4	4
2	4	1	3
4	4	3	5
6	3	2	4
2	4	3	4
2	5	4	4
5	5	1	4
3	5	3	4
3	5	2	5

Example

- Frequencies
  - Image After < Image Before: 3
  - Image After > Image Before: 13
  - Image After = Image Before: 4
  - Total: 20
- One-tailed significance
  - 0.0106

Power of statistical tests

- Statistical tests are only as good as their reliability
- Two types of errors are possible

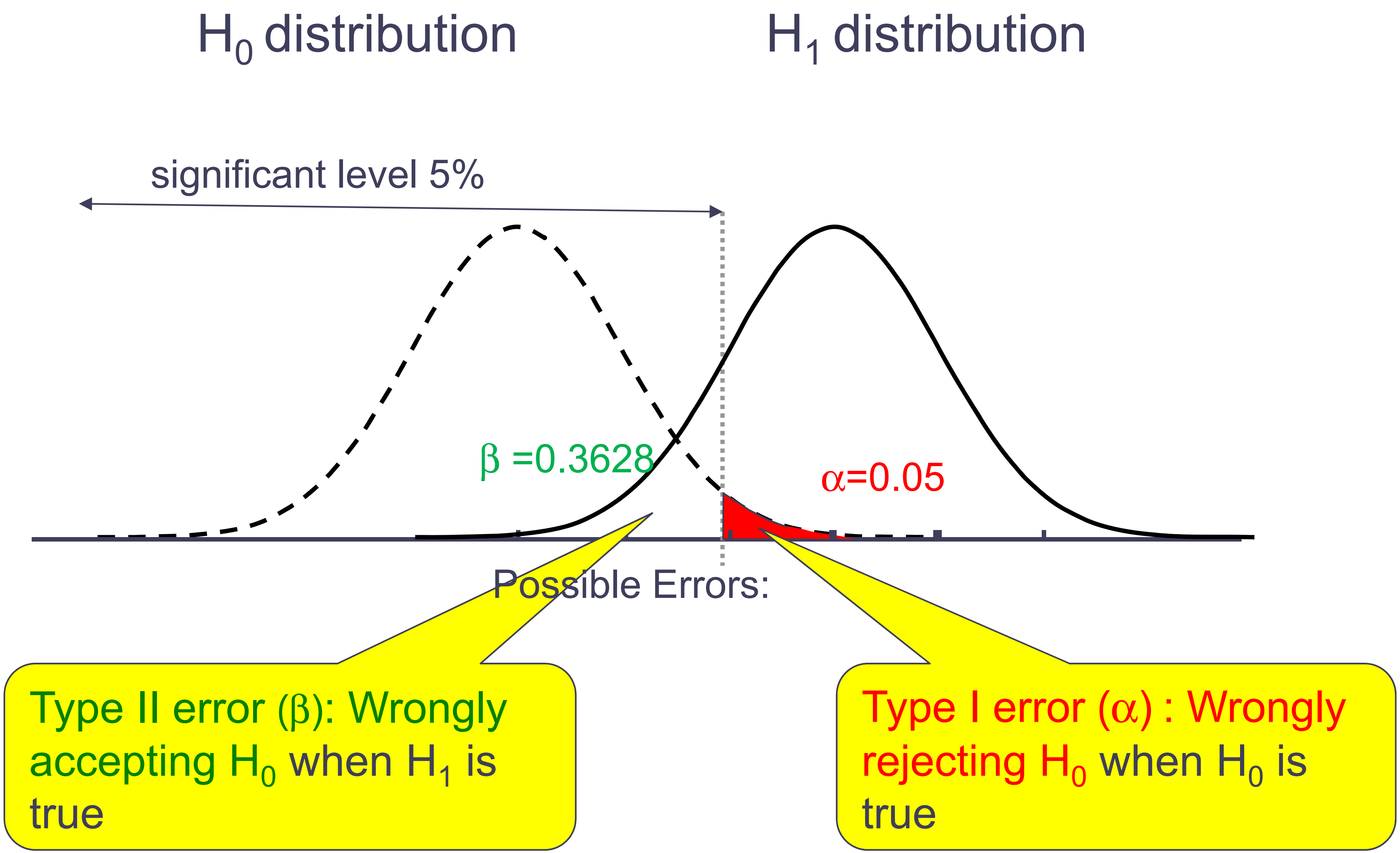
  Type I - Reject H_0 when it is true (α)

  Type II - Retain H_0 when it is false (β)
- Type I & II errors have probabilities denoted α and β
- The power of a statistical test = 1-β,

  - i.e. the probability that it will reject H_0 if H_0 is false

Hypothesis testing		Test result	
		Retain H_0	Reject H_0
H_0	True	1 - α	Type I error α
	False	Type II error β	1- β

Type I and Type II errors



Multiple testing

- If you do many significance tests, then it is likely that you will incorrectly reject the null hypothesis for at least some of them.
- Family-wise error rate (FWER)
  - The risk that any out of a ‘family’ of tests will incorrectly reject the null hypothesis.

Bonferroni correction

- Let n be the number of tests.
- Set α to desired maximum risk of any error.
- Reject all null-hypotheses where $p \leq \alpha/n$.
- Unnecessarily strict
  - Reduces power of individual tests more than necessary to control FWER.

Bonferroni correction example

- p values
  - 0.001, 0.105, 0.021, 0.009, 0.011
- α
  - 0.05
- Bonferroni adjusted $p \leq \alpha/n = 0.05/5$
  - 0.01
- Reject hypotheses with p=0.001 and p=0.009

Holm procedure

- Order *p* values from lowest to highest.
- Assign them indexes *i* from 0 (for lowest *p* value) to *n*-1 (for highest *p* value).
- Test each p_i against $p_i \leq \alpha/(n-i)$ until the first one fails, then reject that hypothesis and all hypotheses with higher *i.*

Holm procedure example

- p values
  - 0.001, 0.105, 0.021, 0.009, 0.011
- Sorted p values
  - 0.001, 0.009, 0.011, 0.026, 0.105
- α
  - 0.05
- Adjusted α values
  - 0.01, 0.0125, 0.0166, 0.025, 0.05
- Reject hypotheses with p=0.001, p=0.009 and p=0.011

Criticisms of hypothesis testing

- "All differences are significant, given large enough sample sizes" (Paul Meehl, 1967, inter alia)
  - Meehl performed a massive survey of HS students in US Midwest (*ca*. 50,000), measuring dozens of attributes. The majority of attributes were correlated to statistical significance (e.g. religion and musical talent).
- "There are lies, damned lies, and then there are statistics"
  - If one test fails, try another that shows what you want!
- Some critics argue that most published studies are wrong
  - Researchers publish only positive results.
  - Negative cases are ignored, so experiments continue until there is a positive result.

Hypothesis testing summary

- Be aware
  - that some significant results will be false positives.
  - if sample size is small, the evidence is weak, check the power.
  - if sample size is huge, the effect size may be small, check the effect size.
  - if H_0 and H_1 are highly divergent, or H_1 is on the other tail, significance may be misleading, (check the likelihood ratio)
- There are many other tests for other cases and for other experimental designs
  - Refer to statistics texts for details.

References

- Leedy, P. D., Ormord, J.E. (2010). *Practical Research: Planning and Design* (8th ed.). Prentice Hall. Upper Saddle River, NJ.
- Siegel, S., Castellan, N.J., Jr. (1989). *Non-parametric Statistics for the Behavioural Sciences.* McGraw-Hill, New York.
- Cramer, D. (1998). *Fundamental Statistics for Social Research: Step by Step Calculations and Computer Techniques Using SPSS for Windows.* Rutledge, London.

Readings

- Web Center for Social Research Methods
  - http://www.socialresearchmethods.net/
  - See section on Selecting Statistics
- Tutorial on tests of significance
  - http://www.csulb.edu/~msaintg/ppa696/696stsig.htm
- McREL (2004) *Tutorial on Understanding Statistics*. ECS.
  - http://www.ecs.org/html/educationIssues/Research/primer/understandingtutorial.asp
  - See section on Selecting Statistics