# Material Summary: Hypothesis Testing

## Confidence Intervals

**1.1 Confidence Intervals**

* In an experiment, we can't observe the variables' true values directly
  + We observe other values
  + We make assumptions as to how they are distributed
  + We can estimate the true value
    - **Law of large numbers**: when our sample is big enough, the sample parameters approach the population parameters
* With continuous values, it's useless to say that the mean is equal to a certain value (why?)
* **Confidence interval** – a range of values that we're fairly sure contains the true value
  + **How confident?** A matter of choice
* **Confidence level** – the probability that the value falls within the interval

**1.2 Confidence Intervals – Interpretation**

* Similar to the probability interpretations
* To illustrate these, let's take a confidence interval and a **70%** confidence level
* Frequency
  + If we perform the experiment many times, 70% of the values will fall in the interval and 30% – outside it
* Certainty of next trial
  + Next time we perform the experiment, we are 70% certain that the value will fall within
  + Note that this is a statement **about the interval**, not about the value
* Typically used confidence levels
  + **50%; 90%; 95%; 99,7%**

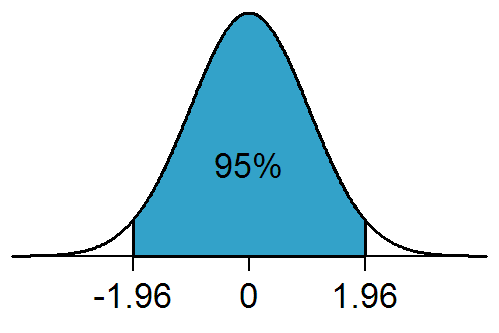
**1.3 Confidence Intervals and Z-Scores**

* Observe the Z-distribution (Gaussian, )
* What's the probability that a value drawn from it ?
  + This corresponds to the shaded area in the graph
  + The cumulative function gives us the area to the left of some value
* Interpretations
  + If we draw many random numbers from the   
    Z-distribution, we expect that 81,9% of them   
    will be in
  + If we draw one random number, there is 81,9%   
    chance of it being in
* Commonly used intervals
  + A blue curve with numbers

    Description automatically generatedAlso

**1.4 Confidence Intervals: Example**

* In the dataset heights.csv you're given the measured heights (in cm) of 351 elderly women (from an osteoporosis study)
  + Plot a histogram and / or boxplot to see what the distribution is
  + Print the mean and standard deviation of the sample
  + Assume that the population follows a normal distribution
    - Real parameters – unknown; our best guess:
  + What are the confidence intervals of
    - 50%, 90%, 95%
* To calculate the confidence intervals, we need to calculate   
  the **Z-scores**
  + To do this, we'll use the percent point function, ppf
    - Inverse of the cdf
    - Returns the value at which the probability is less than or equal to the given probability
    - Example: Z-distribution
* Note that once again we need to subtract the left white region
  + Area of shaded region: (e.g., )
  + Area of both tails:
  + Percentage point of left tail:
  + A screenshot of a computer

    Description automatically generatedPercentage point of right tail:

## Testing Hypotheses

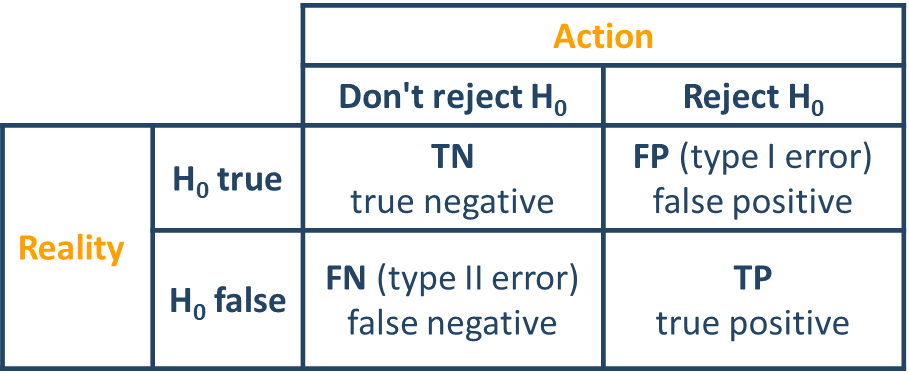
**2.1 Hypotheses**

* After performing an experiment and getting data, the scientific method requires that we form a hypothesis
  + Fact, law, theory and hypothesisare[*different terms*](http://lifehacker.com/the-difference-between-a-fact-hypothesis-theory-and-1732904200)
* In the simplest case, we have two hypotheses
  + **Null hypothesis** () – the status quo is real, "nothing interesting happens"
  + **Alternate hypothesis** () – what we're trying to demonstrate
* Types of hypotheses
  + Attributive – something exists and can be measured
  + Associative – there is a relationship between two behaviors
  + Causal – differences in the amount / kind of one behavior cause differences in other behaviors

**2.2 Hypotheses – Examples**

* Examples of hypotheses – study of Disneyland visitors
  + **Attributive**
    - Most of the population has heard of Disneyland
    - Disneyland visitors are diverse in demographics
  + **Associative**
    - Income level is correlated with visiting Disneyland
    - People who live closer to Disneyland are more apt to visit Disneyland
  + **Causal**
    - Frequent exposure to Disneyland advertising results in increased attendance
    - Discounting tickets for local residents produces an increase in   
      visitor numbers
* Note that attributive hypotheses involve one variable (univariate) while associative and causal hypotheses involve two variables (bivariate)

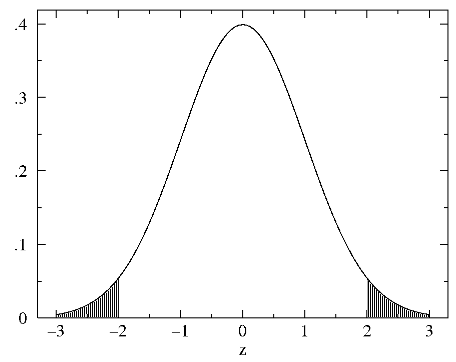
**2.3 Testing a Hypothesis**

* In random experiments, we have error sources
  + Human error, systematic error, random errors, etc.
* We cannot prove (or reject) a hypothesis with complete certainty
* The errors we can make are two types
  + **Type I error** – reject while it's true (false positive)
  + **Type II error** – accept while is true (false negative)
* The possible results can be summarized in the following truth table
  + Also called **confusion matrix**
* To measure the probability of producing a wrong hypothesis, we use a **test statistic** – measure of deviations from
  + Different tests produce different measures (statistics)
  + **We accept or reject the null hypothesis based on the value of the   
    test statistic**
* Let's denote **the probability of getting a type I error** with
  + Each value of the selected test statistic has a corresponding alpha-value
  + We perform the experiment, get data and calculate the test statistic value
  + From that, we calculate the corresponding alpha-value
  + We reject the null hypothesis if , where is a **critical confidence level**

**2.4 Z-test**

* A Z-test uses the Z-statistic
* : standard normal distribution
* Example: light bulb factory
  + A factory produces light bulbs with lifetime
  + A sample of 25 bulbs has a mean lifetime
  + Is there something wrong with the production line?
* Forming hypotheses
  + : The production line works normally; the observed deviation of the sample mean from the population mean is due to chance
  + : The production line is broken
* Suppose we take a lot of samples from the entire population
  + Each sample mean will be different
  + The distribution of sample means will be more or less Gaussian
    - Parameters (our best estimate):
    - [*Here's why*](http://stattrek.com/sampling/sampling-distribution.aspx)the parameters are chosen like this
* If is correct, we assume that
* Z-statistic
* We can see that we are 2 std's below the mean
* How extreme is that?
  + What's the probability that we get results **as extreme or more extreme** than we observed, assuming the null hypothesis is true?
    - Less than 5%

**2.5 Two-tailed Z-test**

* We can get the confidence interval from the Z-statistic
* We are looking for **more extreme** values
  + Values **outside** the confidence interval
  + What's the probability ?
  + We're looking for a value different than the mean
    - We **can't assume** whether it's smaller or larger
    - Therefore, we have to look at both "tails"  
      of the distribution
* If we assume a critical value (also called a p-value) of 5%, **the results   
  are significant**
* We can **reject H0 at the 5% level** 
  + Even at lower levels, up to 4,55%

**2.6 One-tailed Z-test**

* The same logic applies, but now we're looking at one tail only
* Question: Is the lifespan **significantly lower** than it should be?  
  Cutoff point:
  + Answer: Yes, at the given significance level
* Question: Is the lifespan **significantly higher** than it should be?
  + Answer: No, at the given significance level

A black background with a black square

Description automatically generated with medium confidence**2.7 t-test**

* *The Z-test requires that we know the standard deviation of   
  the population*
  + *Usually not available*
* *We can use another test statistic, called* ***t***
* *Advantages over the Z-test*
  + *We don't need to know the population*
  + *It's better when we have very small sample sizes (e.g., )*
  + *It can be used for testing the mean of a sample against a standard, but also, for comparing two means*
    - *We can see whether two sets of data are significantly different from   
      each other*
* ***Null hypothesis****: The test statistic follows Student's t-distribution*
  + *Similar to Gaussian distribution, with "fatter" tails*

**2.8 Moments of the Gaussian Distribution**

* Generalization of the binomial distribution
* A black background with a black square

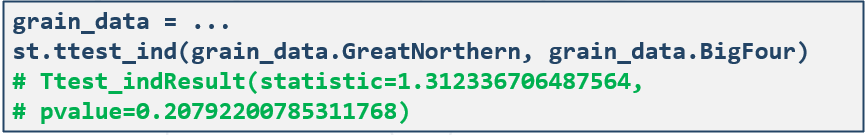
  Description automatically generated with medium confidenceProbability density function
* Mean:
* Median:
* Mode:
* Variance:
* Skewness:
* Excess kurtosis:
* *"And that, kids, is why I love the Gaussian distribution."*

**2.9 One-Sample t-test**

* The details of the calculation are fairly complex but we can do this in code
  + Using scipy.stats
* First, we generate 100 random numbers with
* Then we ask whether the sample mean is equal to the true mean  
  (and other values, just for testing)
* We get the p-value – probability of the null hypothesis being true
  + **A screenshot of a computer code

    Description automatically generated**I.e., probability that the mean is equal to the given mean

**2.10 Independent Two-Sample t-test**

* We compare two independent distributions
  + We want to see whether they have the same mean
  + We assume equal variances (scipy can also do tests   
    with unequal variances – important when sample sizes differ)
* Example: Grain size
  + We are given data (in grain\_data.csv) of grain sizes from two different farms
  + Do they differ significantly (at the 95% level)?
  + ****\* We can also plot histograms to see what the distributions look like

**2.11 Paired Two-Sample t-test**

* We compare two distributions
  + Observations in samples can be paired
  + Examples – before / after observations; comparison between two different treatments applied to the same subjects
* Example: Drinking water
  + We are given data (in water\_data.csv) of Zn concentration in surface and bottom water at 10 different locations
  + Does the true average concentration in bottom water exceed that of top water?
  + We use a paired t-test because the samples are from the same locations
  + **A white rectangular sign with blue text

    Description automatically generated**It reduces experimental error (and provides stronger evidence)

**2.12 Generalizations to More Variables**

* Sometimes it's not enough to compare two distributions
  + We may want to compare multiple distributions against the same null hypothesis
  + E.g., how is the percentage of smokers distributed by income and age?
* Other times, we create a model and want to evaluate it
  + E.g., a linear regression
  + We can explain some of the variance in the sample
* There are other tests to perform these "checks"
  + **ANOVA** (Analysis of Variance) – useful for grouped data
    - Observe the variance inside groups and between groups
  + **Chi-square(d) test** – can be applied to categorical data
    - Two common types
      * How good a model is (goodness of fit)
      * Whether two variables are independent

**2.13 Analysis of Variance (ANOVA)**

* We want to compare several groups
* The means of the groups are the same
* Method (*[scipy.stats.f\_oneway()](https://docs.scipy.org/doc/scipy-0.19.1/reference/generated/scipy.stats.f_oneway.html))*
  + For each group group mean
    - In-group variance: distances from an individual point to the group mean
    - Between-group variance: distances between the means of two groups
  + For the entire data total mean (mean of all data)
    - Also equal to the mean of all group means
    - Total variance: in-group + between-group
* F-statistic (Fisher)
  + - – large the variance between groups dominates
    - For each value of , there’s a corresponding -value
      * If , we can reject

**2.14 Chi-Squared () Test**

* Compares expected (predicted) and observed frequencies
  + Is there a significant difference between these?
  + Used to compare categories (one against another)
    - Compare to ANOVA – numbers w.r.t. categories
  + May also be used as a goodness-of-fit measure
    - How well were we able to predict
* Statistic:
* : No significant difference between observed and estimated frequencies among the categories (groups)
  + The test returns the value of the statistic and the p-value corresponding   
    to it
  + Works the same as any other test
  + Python: [*scipy.stats.chisquare()*](https://docs.scipy.org/doc/scipy-0.19.1/reference/generated/scipy.stats.chisquare.html)

## Common Misconceptions

**3.1 Some p-value Misconceptions**

* Goodman, S. (2008), [*source*](https://www.sciencedirect.com/science/article/abs/pii/S0037196308000620?via%3Dihub)
* "If , has 5% chance of being true"
  + The data alone can't tell us how likely we are to be wrong
  + is calculated under , so it can't be the probability of being false
* " means that if we reject , the probability of type I error (false positive) is only 5%"
  + I.e., seeing a difference where there isn't any
  + 5% chance of false rejection = 5% chance is true
    - Wrong, see first bullet
* "If , we have observed data that will occur only 5% of the time assuming "
  + The p-value is the probability of observing data as extreme or more extreme under
* "A nonsignificant difference means the groups are the same"
  + It only means **we don't have enough data** to reject
* "A scientific conclusion or treatment policy must be based on whether or not the -value is significant"
  + **The results have to be checked** against prior data
* Failing to reject means that is true
  + It means that we don't have enough evidence to reject it
  + **We can't accept (or reject) any other hypothesis**
  + *"Absence of evidence is not evidence of absence"*
* [*https://xkcd.com/882/*](https://xkcd.com/882/)
* [*https://www.xkcd.com/1478/*](https://www.xkcd.com/1478/)
* [*"Still. Not. Significant"*](https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/)article