Material Summary: Hypothesis Testing

1. 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 $[5;\ 7,3]$ and a **70%** confidence level
- Frequency
    - If we perform the experiment many times, 70% of the values will fall in the interval $[5;\ 7,3]$ and 30% – outside it
- Certainty of next trial
    - Next time we perform the experiment, we are 70% certain that the value will fall within $[5;\ 7,3]$
    - 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, $\mu = 0,\ \sigma = 1$)
- What's the probability that a value drawn from it $x \in [-2; 1]$?
    - This corresponds to the shaded area in the graph
    - The cumulative function gives us the area to the left of some value
    - Shaded area $= cdf(1) - cdf(-2) = 0,819 = 81,9\%$
- Interpretations

---

- If we draw many random numbers from the Z-distribution, we expect that 81,9% of them will be in $[-2; 1]$
- If we draw one random number, there is 81,9% chance of it being in $[-2; 1]$

- Commonly used intervals
  - $1\sigma \rightarrow 68,27\%;\ 2\sigma \rightarrow 95,45\%;\ 3\sigma \rightarrow 99,73\%$
  - Also $1,96\sigma \rightarrow 95\%$
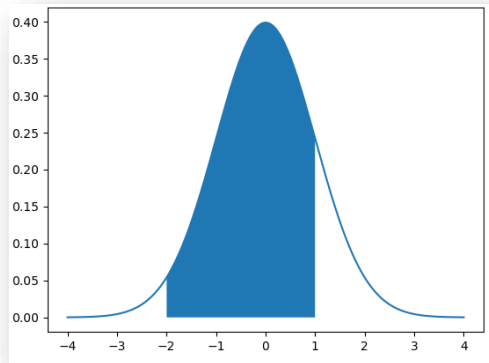
1.4 Confidence Intervals: Example

- 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 $\bar{x}$ and standard deviation s of the sample
  - Assume that the population follows a normal distribution
    - Real parameters – unknown; our best guess: $\mu = \bar{x}$, $\sigma = s$
  - 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
      - $ppf(0) = -\infty;\ ppf(1) = \infty;\ ppf(0,5) = 0;\ ppf(0,975) = 1,96$
- Note that once again we need to subtract the left white region
  - Area of shaded region: p (e.g., $p = 0,95$)

- Area of both tails: $1 - p$

- Percentage point of left tail: $\frac{1-p}{2}$

- Percentage point of right tail: $\frac{1-p}{2} + p = \frac{1-p+2p}{2} = \frac{1+p}{2}$
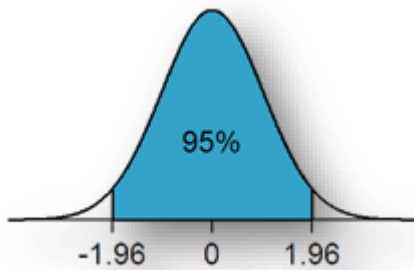

```python
import scipy.stats as st
def get_real_confidence_interval(probability, mean, std):
  lower_area = (1 - probability) / 2
  upper_area = (1 + probability) / 2
  return [
    st.norm.ppf(lower_area, mean, std),
    st.norm.ppf(upper_area, mean, std)]
```

2. 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 hypothesis are *different terms*
- In the simplest case, we have two hypotheses
  - **Null hypothesis** (H_0) – the status quo is real, "nothing interesting happens"
  - **Alternate hypothesis** (H_1) – 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 H_0 while it's true (false positive)
  - **Type II error** – accept H_0 while H_1 is true (false negative)
- The possible results can be summarized in the following truth table
  - Also called **confusion matrix**

		Action	
		Don't reject H_0	**Reject H_0**
Reality	**H_0 true**	**TN** true negative	**FP** (type I error) false positive
	H_0 false	**FN** (type II error) false negative	**TP** true positive

- To measure the probability of producing a wrong hypothesis, we use a **test statistic** – measure of deviations from H_0
  - 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 α
  - 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 $\alpha < \alpha_c$, where α_c is a **critical confidence level**

2.4 Z-test

- A Z-test uses the Z-statistic
- H_0: standard normal distribution
- Example: light bulb factory
  - A factory produces light bulbs with lifetime $X \sim N(\boldsymbol{\mu = 500h,\ \sigma = 50h})$
  - A sample of 25 bulbs has a mean lifetime $\bar{x} = 480h$

- Is there something wrong with the production line?
- Forming hypotheses
    - H_0: The production line works normally; the observed deviation of the sample mean from the population mean is due to chance
    - H_1: 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): $\mu_{\bar{x}} = \mu$, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
        - *Here's why* the parameters are chosen like this
- If H_0 is correct, we assume that $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- Z-statistic
    - $Z = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \dfrac{480-500}{50/\sqrt{25}} = -2$
- 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 $P(|Z| \geq 2)$?
    - 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**
    - $P(|Z| > 2) \approx 0{,}0455 = 4{,}55\%$
- We can **reject H₀ at the 5% level**

- Even at lower levels, up to 4,55%
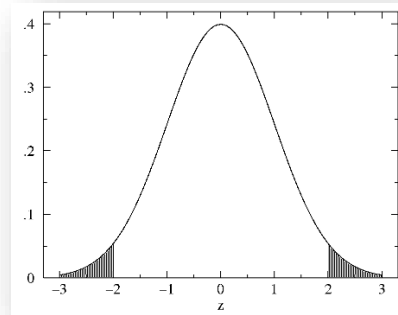
2.6 One-tailed Z-test

- 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: $\alpha_c = 5\%, Z = -2$
  - $P(Z \leq -2) = \frac{0,00455}{2} - 0,02275 = 2,275\% < \alpha_c$
  - Answer: Yes, at the given significance level
- Question: Is the lifespan **significantly higher** than it should be?
  - $P(Z \geq -2) = 97,725\% \gg \alpha_c$
  - Answer: No, at the given significance level

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 σ*
  - *It's better when we have very small sample sizes (e.g., $n < 30$)*
  - *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*

$$\mu_r(x) = \frac{\sum(x - \mu)^r}{N}$$

  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
- Probability density function

$$N(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean: μ
- Median: μ
- Mode: μ
- Variance: σ^2
- Skewness: 0
- Excess kurtosis: 0
- *"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 $\mu = 5, \sigma = 10$
- 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
    - I.e., probability that the mean is equal to the given mean

```python
sample_data = st.norm.rvs(5, 10, 100)

print(st.ttest_1samp(sample_data, 5).pvalue) # 0.9301
print(st.ttest_1samp(sample_data, 4).pvalue) # 0.3352
print(st.ttest_1samp(sample_data, 0).pvalue) # 1.104e-6
```

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

```python
grain_data = ...
st.ttest_ind(grain_data.GreatNorthern, grain_data.BigFour)
# Ttest_indResult(statistic=1.312336706487564,
# pvalue=0.20792200785311768)
```

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
    - It reduces experimental error (and provides stronger evidence)

```
water_data = ...
# We use a one-tailed t-test
st.ttest_rel(water_data.surface, water_data.bottom).pvalue / 2
# 0.00044555772891127738
```

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
- H_0: The means of the groups are the same
- Method (*scipy.stats.f_oneway()*)
    - For each group $\Rightarrow$ 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)
  - $F = \dfrac{\text{variance between groups}}{\text{variance within groups}}$
    - F – large ⇒ the variance between groups dominates
    - For each value of F, there's a corresponding p-value
      - If $p \leq p_c$, we can reject H_0

2.14 Chi-Squared (χ^2) 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: $\chi^2 = \dfrac{(f_{\text{observed}} - f_{\text{estimated}})^2}{f_{\text{estimated}}}$
- H_0: 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()*

3. Common Misconceptions

3.1 Some p-value Misconceptions

- Goodman, S. (2008), *source*
- "If $p = 0,05$, $H0$ has 5% chance of being true"
  - The data alone can't tell us how likely we are to be wrong
  - p is calculated under $H0$, so it can't be the probability of $H0$ being false
- "$p = 0,05$ means that if we reject $H0$, the probability of type I error (false positive) is only 5%"
  - I.e., seeing a difference where there isn't any

- - $\Rightarrow$ 5% chance of false rejection = 5% chance $H0$ is true
    - Wrong, see first bullet
- "If $p = 0,05$, we have observed data that will occur only 5% of the time assuming $H0$"
    - The p-value is the probability of observing data as extreme or more extreme under $H0$
- "A nonsignificant difference means the groups are the same"
    - It only means **we don't have enough data** to reject $H0$
- "A scientific conclusion or treatment policy must be based on whether or not the p-value is significant"
    - **The results have to be checked** against prior data
- Failing to reject $H0$ means that $H0$ 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://www.xkcd.com/1478/*
- *"Still. Not. Significant"* article