[[Hypothesentests]]

- Basic idea: probabilistic "proof" by contradiction
  1. We assume that a hypothesis H holds
  2. We compute how likely is our data if H holds
  3. If the data is not "very" likely we reject H

Z-Test

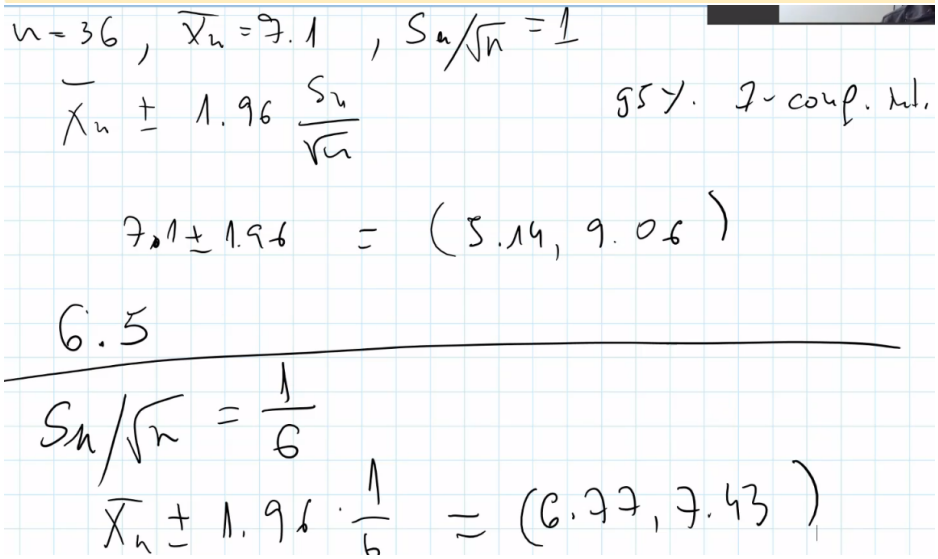
- Let $X_1, X_2, \ldots, X_n$ be a random sample of size n with $X_1 \sim N(\mu, \sigma^2)$, μ unknown, σ known.

Hypotheses:	$H_0 : X_1 \sim N(\mu_0, \sigma^2)$
	$H_1 : \mu \neq \mu_0$ (two sided alternative)
	or $H_1 : \mu > \mu_0$ (one sided alternative)
Test statistic:	Z_n
Null distribution:	Assuming H_0: $Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
Significance level:	α
Critical value:	$c = z_{1-\alpha/2}$ (two sided alternative)
	$c = z_{1-\alpha}$ (one sided alternative)
Rejection region:	$R = \{\lvert Z_n \rvert > z_{1-\alpha/2}\}$ (two sided alternative)
	$R = \{Z_n > z_{1-\alpha}\}$ (one sided alternative)

Examples

- user satisfaction
  - We perform a user study including $n = 36$ users to estimate the user satisfaction with the new app version. The average user satisfaction in our user study is 7.1. The current version of the app is in use for a couple of years now, and we know from numerous user studies and online rating sites that the average user satisfaction with the current version is about 6.5. The statistical model here is: we are dealing with **two populations** of users, those using the current version and those using the new version of the app. Traditionally, to answer this type of the question we set up a **test to check if there is a substantial evidence that one mean is greater than the other mean**.
  - We formulate the hypothesis that the new version is no better than the current version of the app. Generally, we hope that with our test we can reject this hypothesis. There is an indication that the users of the new version are more satisfied, but as previously discussed, **the sample mean is not sufficient, and we need to estimate the sample variability**.

Suppose now that the sample deviation $S_n/\sqrt{n}$ in our user study with the new version of the app is 1. Then a 95% Z-score confidence interval (assuming normality of the user population of the new version of the app) is $(5.14, 9.06)$. Thus, the sample mean of 7.1 could easily have from a population with mean smaller than 6.5 (user satisfaction average with the current app version). Thus, we have no strong ground for rejecting the hypothesis. If, on the other hand $S_n/\sqrt{n}$ were 0.167, then Z-score confidence intervals would be $(6.77, 7.43)$ and we could confidently reject the hypothesis and pronounce the new version of the app to be superior with respect to user satisfaction.



- possible test

  Let $X_1, X_2, \ldots, X_n$ be a random sample of size n with $X_1 \sim N(\theta, 25)$, where θ is an unknown population mean. Consider $H : \theta \leq 17$. One possible test is as follows: reject H if and only if $\overline{X}_n > 17 + 5/\sqrt{n}$.

- unfair coin

  We toss a coin $n = 80$ times and get heads 54 times. Can we conclude that the coin is significantly unfair?

  With our data we obtain:

$$\frac{0.675 - 0.5}{\sqrt{0.5(1-0.5)}/\sqrt{80}} = 3.1305$$

  3.1305 is not a plausible realization of a r.v. $Z \sim N(0,1)$ because:

$$P(Z > 3.1305) = 1 - \Phi(3.1305) = 0.0009$$

  Conclusion: It **seems quite reasonable** to reject H.

With $\overline{X}_n = 13/30 \approx 0.433$:

$$\frac{0.433 - 0.5}{\sqrt{0.5(1-0.5)}/\sqrt{30}} = -0.7303$$

-0.7303 is a plausible realization of a r.v. $Z \sim N(0,1)$, e.g.

$$P(Z > -0.7303) = 1 - \Phi(-0.7303) = 0.7674$$

Conclusion: Our data **does not suggest** to reject H.

- First user group: users of the new app version
- Second user group: users of the current app version
- Typically we will formulate **two mutually exclusive hypotheses:**
  1. **The Null Hypothesis** H_0**:** The user satisfaction is the same in the two groups.
  2. **The Alternative Hypothesis** H_1**:** The user satisfaction is not the same in the two groups.

Remark 11 (The Null vs. The Alternative)

H_0 is always **cautious default**: we won't claim claim the coin is unfair unless we have a strong evidence! You can also think about H_0 as "nothing interesting is happening" and of H_1 as "something interesting is going on". For example, in a legal trial we always assume someone is innocent (H_0) unless the evidence strongly suggest that the person is guilty (H_1).

- unfair coin 2

In the case of coin tosses we test:

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

H_0: The coin is C_2 vs. H_1: The coin is C_1. $\alpha = 0.05$
$0.5 < 0.6 \implies$ one sided (left) rejection region

k	0	1	2	3	4	5	6	7	8
$p(k\|p=0.5)$	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
$p(k\|p=0.6)$	0.001	0.008	0.041	0.124	0.232	0.279	0.209	0.090	0.017

Since we got 6 heads we do not reject H_0.
Paradox: the fact that we don't reject C_1 in favor of C_2 or C_2 in favor of C_1 reflects asymmetry of hypothesis testing. The null hypothesis is a cautious choice. We only reject H_0 if the data is extremely unlikely when we assume H_0. This is not the case for either C_1 or C_2.

- radar guns

Suppose $n = 3$ radar guns are set up along a stretch of road to catch people driving over the speed limit of 50 km/h. Each radar gun is known to have a normal measurement error $N(0, 5^2)$. For a passing car, let $\overline{X}_n$ be the average of the n readings. The police's default assumption is that car is not speeding (cautious choice: innocent until proven guilty!).

1. Describe the above story in the context of statistical hypothesis testing. Are the most natural null and alternative hypotheses simple or compound?
2. The police would like to set a threshold on $\overline{X}_n$ for issuing tickets so that no more than 4% of the tickets are given in error. Use the statistical hypothesis testing described in 1 to determine what threshold the police should set when using $n = 3$ radars.

The null distribution is $\overline{X}_n \sim N(50, 5^2/3)$ and we are looking for one sided rejection region $(H_1\colon\ \mu > 50)$ with $\alpha = 0.04$:

$$c = \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 50 + 1.7507\frac{5}{\sqrt{3}} = 55.054 \text{ km/h}$$

Simple vs. composite hypotheses

Let μ be the actual speed of a given driver. Hence, $X_1 \sim N(\mu, 5^2)$ and therefore $\overline{X}_n \sim N(\mu, 5^2/3)$. The most natural hypothesis are:

1. $H_0\colon \mu \leq 50$, i.e., the driver is not speeding
2. $H_1\colon \mu > 50$, i.e., the driver is speeding

Both hypotheses are composite, however we can work with $H_0\colon \mu = 50$, which is simple.

Critical value for rejecting $H_0\colon \mu \leq 50$

The null distribution is $\overline{X}_n \sim N(50, 5^2/3)$ and we are looking for one sided rejection region $(H_1\colon \mu > 50)$ with $\alpha = 0.04$:

$c = \mu + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

We compute this value with scipy.stats library, using the normal distribution and the percentile function **ppf**.

```
c = stats.norm.ppf(1 - alpha, mu, sigma / math.sqrt(n))
print('Critical value for rejecting H_0: c = {:6.4f} km/h'.format(c))
```

```
Critical value for rejecting H_0: c = 55.0538 km/h
```