Probability and Statistics for Data Science

Data Science Course

Dr. Ariel Mantzura

2025-02-06

Topics to be Covered in this Lecture

- Statistical Inference - Introduction
- IID Random Samples
- Limits of Random Variables
- The Law of Large numbers
- The Central limit theorem
- Confidence Intervals
- Hypotheses testing
- Point estimation
    - Method of moments
    - Maximum Likelihood
- Unbiasedness

Source: **Statistical Inference for Data Science** Brian Caffo

Statistical Inference - Introduction

- Statistical inference is the process of analyzing data to infer properties of an underlying data generating process.
- It is assumed that the observed data set is sampled from some unknown data generating process.
- Without statistical inference we're simply living within our data.
- With statistical inference, we're trying to generate new knowledge about the underlying mechanism that generated the data.

Statistical Inference - Introduction

- Any statistical inference requires some assumptions (strong or weak) regarding the generation of the observed data and similar data.
- These assumptions usually focus on some (DGP) quantities of interest, about which we wish to draw inference.
- One simple example is: we may assume that some data was generated from a normal distribution where the parameters of interest, i.e. μ and σ are unknown.

Statistical Inference - Introduction

- Some common types of statistical inference are the following:
  - **A point estimate**, i.e. a particular value that best approximates some parameter of interest. Example: a point estimation of the mean given some data.
  - **An interval estimate**, e.g. a confidence interval (or set estimate), i.e. an interval constructed using a dataset drawn from a DGP so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level.
  - **Hypotheses testing** acceptance or rejection of a hypothesis regarding the parameters of interest. Example: Is the mean of a normal distribution larger that 15 or not based on observed data.

Statistical Inference - Introduction

- Statisticians distinguish between two levels of DGP assumptions:
  - **Fully parametric**: The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. For example, one may assume that the DGP is truly Normal, with unknown mean and variance,
  - **Non-parametric**: The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal. For example, every continuous probability distribution has a median, which may be estimated using the sample median.

I.I.d Random Samples - The Basic assumption

- We've learned about random variables and independence.
- We can introduce a central modeling assumption made in statistics.
- Specifically the idea of a random sample.
- Random samples are said to be independent and identically distributed (iid) if they are independent and all are drawn from the same DGP.
- This is a default starting point for most statistical inferences.

I.I.d Random Samples - examples

- $x_1, x_2, x_3, \ldots, x_n$ are all independent draws from the Binomial distribution $Binom(10, 0.3)$
- $x_1, x_2, x_3, \ldots, x_n$ are all independent draws from the Poisson distribution $Poiss(2)$
- $x_1, x_2, x_3, \ldots, x_n$ are all independent draws from the Normal distribution $N(2, 3)$
- $x_1, x_2, x_3, \ldots, x_n$ are all independent draws from the Exponential distribution $exp(3)$
- $x_1, x_2, x_3, \ldots, x_n$ are all independent draws from the same unknown distribution.

Limits of random variables

- We will discuss the limiting behaviour of one statistic: **The sample average**.
- By limiting behavior we mean the following: If we draw a sample:

$$x_1, x_2, \ldots, x_n$$

  that are assumed to be independent draws from the same DGP.
- We then calculate $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- We may ask what are the properties of $\bar{x}$?
- How close is it to the true mean of the DGP?
- Can we say anything about $\bar{x}$ distribution?

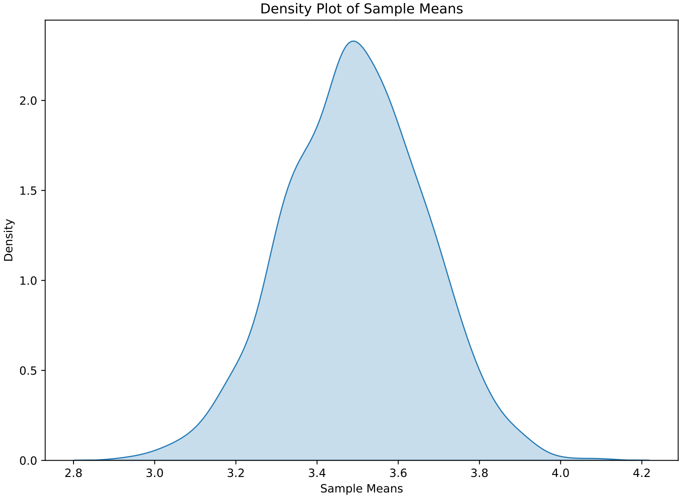
Example 1 - Tossing a fair die

- What is the mean of a faie die toss?
- Toss a fair die n=10 times, n=100 times, n=1000 times.
- Calculate the sample mean(average) for each n?
- Plot the density of 1000 replications when n=100, i.e. a thousand times tossing a fair die 100 times and for each 100 times calculate the average so the we have 1000 averages.

Example 1 - Tossing a fair die

```
## 10 samples: [3.7, 2.8, 4.5, 3.0, 3.2]

## 100 samples: [3.71, 3.78, 3.36, 3.83, 3.16]

## 1000 samples: [3.493, 3.452, 3.401, 3.558, 3.491]

## 10000 samples: [3.505, 3.517, 3.5155, 3.4945, 3.5128]
```

Example 1 - Tossing a fair die



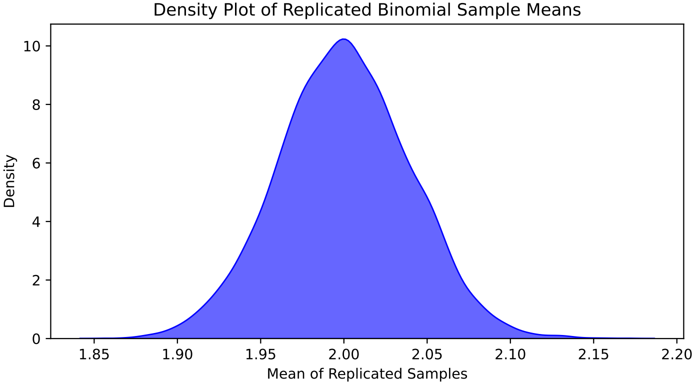
Example 2 - generating i.i.d random draws from Binomial distribution.

- What is the mean of a $X \sim Bin(10, 0.2)$?
- Randomize a binomial(10,0.2) 100 times, 1000 times, 10000 times.
- Calculate the sample mean(average) for each size?
- Plot the density of 10000 replications when size=100, i.e. a thousand times randomize a binomial(10,0.2) 100 times and for each 100 times calculate the average so the we have 1000 averages.

Example 2 - generating i.i.d random draws from Binomial distribution.

```
## n=10: [2.5, 2.5, 2.3, 1.8, 2.3]

## n=100: [1.98, 1.87, 1.95, 2.01, 2.07]

## n=1000: [2.007, 1.997, 2.001, 1.972, 1.934]

## n=10000: [2.0393, 1.9792, 1.9963, 1.9859, 1.9853]
```

Example 2 - generating i.i.d random draws from Binomial distribution.



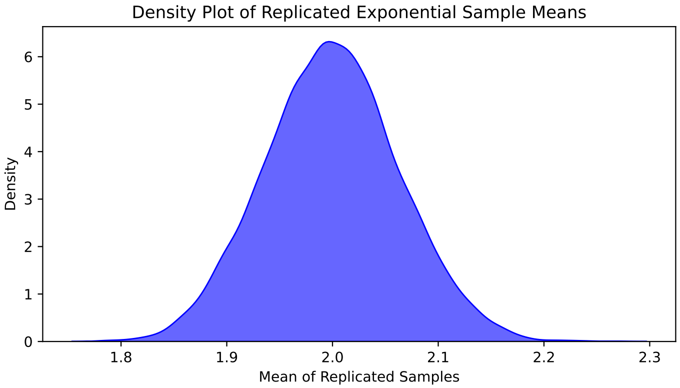
Example 3 - generating i.i.d random draws from Exponential distribution.

- What is the mean of a $X \sim exp(2)$?
- Randomize a exp(2) n=100 times, n=1000 times, n=10000 times.
- Calculate the sample mean(average) for each size?
- Plot the density of 10000 replications when size=100, i.e. a thousand times randomize a exp(2) 100 times and for each 100 times calculate the average so the we have 1000 averages.

Example 3 - generating i.i.d random draws from Exponential distribution.

```
## n=10: [2.568 2.019 1.298 2.292 1.92 ]

## n=100: [1.996 1.728 2.096 1.904 2.478]

## n=1000: [2.085 1.995 1.951 2.013 1.944]

## n=10000: [1.996 1.997 2.01  2.001 2.043]
```

Example 3 - generating i.i.d random draws from Exponential distribution.



Law of Large Numbers

- The (Weak) law of large numbers states that the average of the results obtained from a large number of independent and identical random samples converge to the true value.
- More formally, the LLN states that given a sample of independent and identically distributed values the sample mean (average) converges to the true mean.
- Mathematically this is written as follows:

$$\bar{X}_n \xrightarrow{P} E(X)$$

- The convergence is in probability (beyond the scope of the course)

Central Limit Theorem (CLT)

- The Central Limit Theorem (CLT) is one of the most important theorems in statistics.
- The CLT states that the distribution of averages of iid variables becomes that of a normal variable as the sample size increases.
- Mathematically this is written as follows:
- Let $X_1, X_2, \ldots X_n$ be an i.i.d random sample and let $E(X) = \mu$ and $Var(X) = \sigma^2$ then:

$$\bar{X}_n \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$$

- Equivalently this can be written:

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

Building confidence interval for the true mean

- Confidence intervals are methods for quantifying uncertainty in our estimates.
- The meaning of the interval is that with high probability (typically) 0.95 the true mean will fall within the interval.
- When we say with high probability we mean that if we take many samples from the same GDP the sample average will fall within this interval 95% of the times.

Building confidence interval for the true mean

- The confidence interval using the CLT is constructed as follows:
- According to the CLT, the sample mean, $\bar{X}$, is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$:

$$\bar{X}_n \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$$

- The confidence interval of 0.95 is defined as:

$$[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}]$$

Building confidence interval for the true mean

- This can be explained by the following argument:
- $$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$
- For a standard Normal random variable Z:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

```
from scipy.stats import norm
probability = norm.cdf(1.96) - norm.cdf(-1.96)
print(probability)
```

```
## 0.950004209703559
```

Building confidence interval for the true mean

- So if we set $Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ we get that:

$$P(-1.96 \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95$$

- Rearranging a bit the terms we get that:

$$P(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$$

Building confidence interval for the true mean

- In the following equation we take the number 1.96 indicating that we take 1.96 standard deviations around the sampleaverage.

$$P(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$$

- In order to be more generic we can take the value $Z_{(1-\frac{\alpha}{2})}$ and in our case $Z_{0.975}$ where $\alpha = 0.025$. So we get

$$P(\bar{X} - Z_{0.975} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{0.975} \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$$

```
z_score = norm.ppf(0.975)
print(z_score)
```

```
## 1.959963984540054
```

Building confidence interval for the true mean - unknown σ

- The previous result is true in case the true σ is known.
- When σ is not known as usually happens we estimate σ by the sample standard deviation:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

.

- In this case we replace the value $Z_{1-\frac{\alpha}{2}}$ by $t_{(n-1,1-\frac{\alpha}{2})}$ so that we get the following equation:

$$P(\bar{X} - t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}) = 1 - \alpha$$

```
from scipy.stats import t
t.ppf(0.975, df=100)
```

```
## 1.9839715184496334
```

Building confidence interval for the true mean - Exercise

- To understand the idea of a confidence interval we will conduct the following rejection and acceptance simulation:
- Define a matrix of dimension 100×1000 where 100 is the sample size and 1000 is the number of replications.
- In each column simulate 100 exponential random variables with $\lambda = 2$.
- Calculate the average of each column.
- Build a confidence interval where $\mu = \frac{1}{2}$ and $\sigma = \frac{1}{2}$.
- Recall that for $X \sim exp(\lambda)$ $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$.
- Count the number of times that the true mean falls in the confidence interval.

Building confidence interval for the true mean - Exercise

```
import numpy as np
np.random.seed(0)
samples=np.random.exponential(scale=1/2,size=(100, 1000))
means = samples.mean(axis=0)
lower_bound = 0.5 - 1.96 * (0.5 / 10)
upper_bound = 0.5 + 1.96 * (0.5 / 10)
rej = (means >= lower_bound) & (means <= upper_bound)

proportion_rej = np.mean(rej)
print(proportion_rej)
```

```
## 0.946
```

Hypotheses Testing - motivation

- Classical hypothesis testing is concerned with deciding between two hypothesese regarding the data generating process of the observed data.
- The first, a null hypothesis is specified that represents the status quo.
- This hypothesis is usually labeled, H_0. This is what we assume by default and called the null hypothesis.
- The alternative or research hypothesis is what we require evidence to conclude.
- This hypothesis is usually labeled, H_a or sometimes H_1.

Hypotheses Testing - Types of errors

- In hypothesis testing there are 4 possible results:
  - H_0 is correct and we decided H_0 - Correctly accepted the null hypothesis.
  - H_0 is correct and we decided H_1 - **Type I error**
  - H_1 is correct and we decided H_1 - Correctly accepted the alternative.
  - H_1 is correct and we decided H_0 - **Type II error**

Hypotheses Testing - 3 Hypotheses types

- There are 3 types of hypotheses for the mean:
- One sided right:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

- One sided left:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

- Two sided test:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Hypotheses Testing - The experiment and test

- We defined the Type I error as the situation where H_0 is true and we rejected H_0, i.e. accepted H_1
- We wish to confine the probability of this error denoted by α to be small.
- Typically $\alpha = 0.05$
- We will write this formally:

$$P(\text{rejecting} \quad H_0 | H_0 \quad \text{is true}) = \alpha$$

Hypotheses Testing - Right sided experiment and test

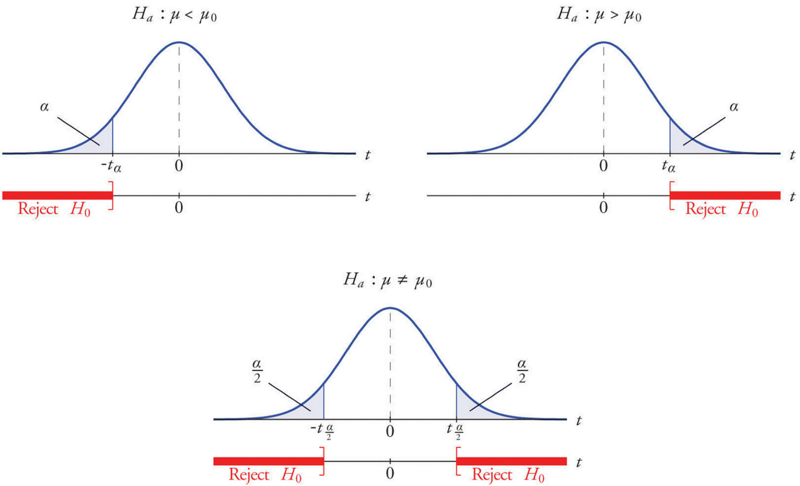
- The experiment: we draw a sample of n observations.
- We then calculate the average.
- We determine a threshold c as follows:

$$P(\bar{X} > c | H_0) = \alpha = 0.05$$

- According to the CLT and assuming H_0 is true: $\bar{X} \sim n(\mu_0, \frac{\sigma^2}{n})$
- So after standardizing $\bar{X}$ and c we get:

$$P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}} | H_0\right) = \alpha = 0.05$$

Hypotheses Testing - The experiment and test



Hypotheses Testing - Right sided experiment and test

- We get that

$$P\left(Z > \frac{c-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - P\left(Z \leq \frac{c-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) =$$

$$= 1 - \Phi\left(\frac{c-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha = 0.05$$

- Rearranging terms we get:

$$\Phi\left(\frac{c-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \alpha = 1 - 0.05 = 0.95$$

- Z is the inverse function of Φ so that:

$$\frac{c-\mu_0}{\frac{\sigma}{\sqrt{n}}} = Z_{0.95} \Rightarrow c = \mu_0 + Z_{0.95}\frac{\sigma}{\sqrt{n}}$$

- We then compare $\bar{x}$ to the threshold c and reject H_0 if:

$$\bar{x} > c$$

Hypotheses Testing

- Right side test the threshold is:

$$c = \mu_0 + Z_{0.95}\frac{\sigma}{\sqrt{n}}$$

- We reject H_0 if

$$\bar{x} > c$$

.

- Left side test the threshold is:

$$c = \mu_0 - Z_{0.95}\frac{\sigma}{\sqrt{n}}$$

- and we reject H_0 if

$$\bar{x} < c$$

.

- In a two side test the thresholds are:

$$c_1 = \mu_0 - Z_{0.975}\frac{\sigma}{\sqrt{n}} \quad c_2 = \mu_0 + Z_{0.975}\frac{\sigma}{\sqrt{n}}$$

Hypotheses Testing - Unnown σ

- As in case of the confidence interval when $\sigma$is not known it is estimated by the sample standard deviation S
- We then take $t_{n-1,1-\alpha}$ instead of $Z_{1-\alpha}$
- The right side test threshold is:

$$c = \mu_0 + t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}$$

- Left side test the threshold is:

$$c = \mu_0 - t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}$$

In a two side test the thresholds are:

$$c_1 = \mu_0 - t_{n-1,1-\alpha}\frac{S}{\sqrt{n}} \quad c_2 = \mu_0 + t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}$$

and

Exercise

- Draw a sample $x_1, x_2, \ldots x_{100}$ from a Binomial(10, 0.2). So that $\mu = np = 2$
- We want to test the following hypotheses:

$$H_0 : \mu = 2$$

$$H_1 : \mu > 2$$

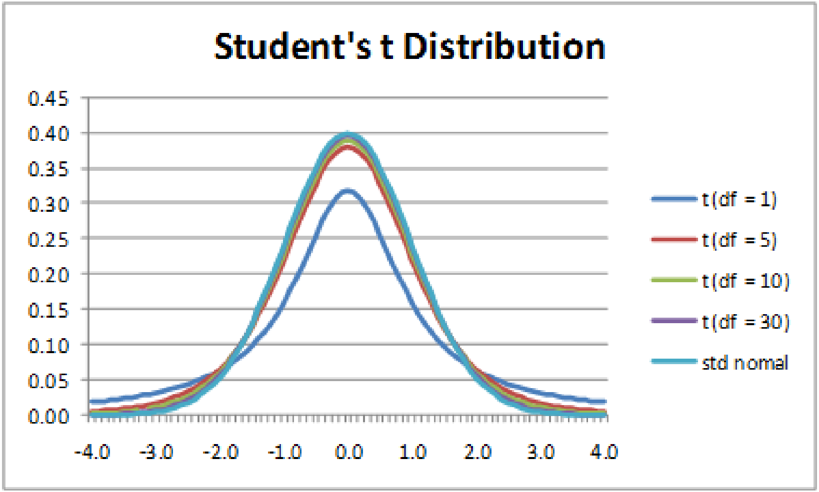
- What should be c.
- Conduct the test on $\bar{x}$
- Draw 1000 samples and conduct the test on each of the samples?
- How many times was H_0 rejected?

Exercise

```
#np.random.seed(0)
rs=np.random.binomial(n=10,p=0.2,size=(100,1000))

# Calculate column-wise means
aves = rs.mean(axis=0)

# Calculate the critical value 'c'
cval=2+((np.sqrt(10*0.2*0.8)/np.sqrt(100))*norm.ppf(0.95))

# Calculate the proportion of means greater than 'c'
proportion_greater = np.mean(aves > cval)

print(proportion_greater)
```

```
## 0.051
```

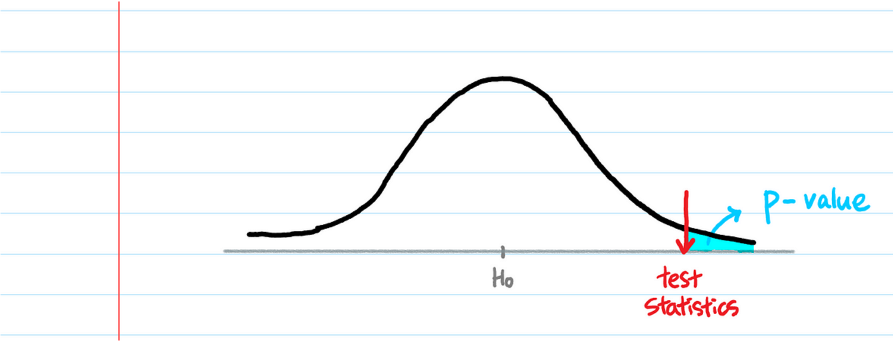
Hypotheses Testing - t distribution



Hypotheses Testing - p values

- We can conduct the hypotheses testing in 2 equivalent methods.
- **Method 1**: Determine the threshold c or thresholds c_1, c_2 for some specified α in a two sided test and examine where $\bar{x}$ fell relative to these thresholds.
- **Method 2**: Calculate $\bar{x}$ and then ask the following question: What should have been the minimal α so that $\bar{x}$ is in the rejection area.
- We then compare the p value with the required significance level.
- If the p value is smaller we reject H_0.

Hypotheses Testing - p value



Hypotheses Testing - p values

- Right sided test:

$$\text{p value} = P(Z > \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

- Left sided test:

$$\text{p value} = P(Z \leq \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}) = \Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

- Two sided test:

$$\text{p value} = \min\left(\Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right), 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)\right)$$

Hypotheses Testing - example

- A machine is used to fill bags with coffee and each bag should contain 1 kg.
- A randomly selected sample of 30 bags has a mean weight of 1.01 kg with a standard deviation of 0.02.
- Perform a two sided statistical test with significance $\alpha = 0.05$ and infer if the machine needs to be adjusted.
- Use both methods to test the hypotheses.

Hypotheses Testing - example

```python
# Calculate a and b
z = (1.01 - 1) / (0.02 / (30**0.5))
a = norm.cdf(z)
b = 1 - norm.cdf(z)
# Print a and b
print(a, b)
```

```
## 0.9969150503397279 0.0030849496602720627
```

```python
# Calculate the p-value
pv = min(a, b)
# Decision
decision = "reject" if pv < 0.025 else "accept"
print(decision)
```

```
## reject
```

Hypotheses Testing - Exercise

- Simulate 1000 samples of size 100 from the $U(0,3)$ distribution.
- What is the mean $E(X)$ of $U(0,3)$? What is the variance and sd?
- Test the following two sided hypotheses.

$$H_0 : \mu_0 = 1.5$$

$$H_1 : \mu_0 \neq 1.5$$

- Use both methods.
- Calculate the proportion of times that H_0 was rejected.

Hypotheses Testing - Exercise

- The mean value is $\frac{0+3}{2} = 1.5$ and the variance is $\frac{(3-0)^2}{12} = 0.75$
- $c_1 = 1.5 - Z_{0.975}\frac{\sqrt{0.75}}{\sqrt{100}} \quad c_2 = 1.5 + Z_{0.975}\frac{\sqrt{0.75}}{\sqrt{100}}$

Hypotheses Testing - Exercise

```
np.random.seed(1)
samps = np.random.uniform(0, 3, size=(100, 1000))
means = np.mean(samps, axis=0)
c1=1.5-norm.ppf(0.975)*(np.sqrt(0.75)/ np.sqrt(100))
c2 =1.5+norm.ppf(0.975)*(np.sqrt(0.75)/np.sqrt(100))
res1=(means > c1) & (means < c2)
zscore=(means-1.5)/(np.sqrt(0.75)/np.sqrt(100))
pv=np.column_stack((norm.cdf(zscore),1-norm.cdf(zscore)))
pv2 = np.min(pv, axis=1)
result1 = np.mean(res1)
result2 = 1 - np.mean(pv2 <= 0.025)
print(result1, result2)
```

```
## 0.954 0.954
```

Point estimation

- Point estimation involves the use of sample data to calculate a single value which is to serve as a "best guess" or "best estimate" of some unknown DGP parameter.
- Point estimation can be contrasted with interval estimation such as confidence intervals which we learned.
- A point estimator can also be contrasted with a distribution estimator (beyond the scope of the course).

Point estimation - Examples

- Assuming $x_1, x_2, \ldots, x_n$ are draws from the **Exponential** distribution we wish to give a point estimate of λ.
- Assuming $x_1, x_2, \ldots, x_n$ are draws from the **Binomial** distribution we wish to give a point estimate of p.
- Assuming $x_1, x_2, \ldots, x_n$ are draws from the **Normal** distribution we wish to give a point estimates of μ and σ.
- Assuming $x_1, x_2, \ldots, x_n$ are draws from the **Poisson** distribution we wish to give a point estimate of λ.

Point Estimation - Method of Moments

- The method of moments is a very simple method for estimating paramaters of importance.
- The basic idea is simply to compare the theoretical moment to the sample moment and accordingly derive the Estimator.

Point Estimation - Method of Moments

- The k^{th} theoretical moment of a random variable is
$$E(X^k)$$
.
- The sample k^{th} moment is defined as
$$\frac{\sum_{i=1}^{n}(x_i^k)}{n}$$
.
- For example the first moment of a random variable is simply its mean:
$$E(X^1) = E(X)$$
.
- The first sample moment is simply the sample average:
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
.

Method of Moments - Example 1

- Assume that $x_1, x_2, \ldots, x_n$ are independent draws from the exponential distribution $X_1, X_2, \ldots, X_n \sim exp(\lambda)$.
- We wish to estimate the rate parameter λ so we will compare the first theoretical moment with the first sample moment.
- The first theoretical moment is $E(X) = \frac{1}{\lambda}$.
- We then compare the following

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{\lambda} = E(x)$$

.

- After a bit manipulating we get that

$$\hat{\lambda} = \frac{1}{\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)} = \frac{1}{\bar{x}}$$

Method of Moments - Example 2

- Assume that $x_1, x_2, \ldots, x_n$ are independent draws from the Binomial distribution $X_1, X_2, \ldots, X_n \sim Binomial(n, p)$.
- We wish to estimate the parameter p.
- The first theoretical moment is

$$E(x) = np.$$

- We then compare the following

$$\frac{\sum_{i=1}^{n} x_i}{n} = np = E(x)$$

.

- After a bit manipulating we get that:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n^2}$$

Method of Moments - Example 3

- Assume that $x_1, x_2, \ldots, x_n$ are independent draws from the Binomial distribution $X_1, X_2, \ldots, X_n \sim Unif(0, \theta)$.
- We wish to estimate the parameter θ.
- The first moment is

$$E(X) = \frac{0 + \theta}{2}.$$

- We then compare the following

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\theta}{2} = E(x)$$

.

- After a bit manipulating we get that:

$$\hat{\theta} = 2\frac{\sum_{i=1}^{n} x_i}{n} = 2\bar{x}$$

Method of Moments - Example 4

- Assume that $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma) \quad (i.i.d.)$.
- We wish to estimate the parameter μ, σ.
- We then compare the following: $\bar{x} = \mu = E(x)$.
- The second moment is $E(X^2) = Var(X) + (E(X)^2) = \sigma^2 + \mu^2$
- We then compare the following:

$$\sigma^2 + \mu^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

.

- Solving for σ^2 and plugging in $\bar{x} = \hat{\mu}$we get:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Point estimation - Maximum Likelihood

- You assume a DGP that depends on some parameter θ (for example, Bernoulli with unknown parameter p), and receive i.i.d. samples $x_1, x_2, \ldots, x_n \sim Ber(p)$ (in this example, each x_i is either 1 or 0).
- The likelihood of the data given a parameter θ is defined as the probability of seeing the data, given the parameter θ, or:

$$L(x|\theta) = P_X(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta)$$

and in our example:

$$L(x|p) = P_X(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|p)$$

Maximum Likelihood - Example

- Suppose we have a sample of the following 4 i.i.d Bernouli realizations:

$$x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 1$$

- We ask what is the p that maximizes the probability to observe these realizations.
- This is the same as asking: what is the p that maximizes thae following probability:

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1|p)$$

.

Maximum Likelihood - Example

- Because of the independence:

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 1|p) =$$

$$= P(X_1 = 1|p)P(X_2 = 1|p)P(X_3 = 0|p)P(X_4 = 1|p)$$

- Because of identically distributed:

$$pp(1-p)p = p^3(1-p) = p^3 - p^4$$

- Taking the first derivative and comparing to 0 we get:

$$3p^2 - 4p^3 = 0 \Rightarrow \hat{p} = \frac{3}{4}$$

Maximum Likelihood - Definition

- Let $x = (x_1, x_2, \ldots, x_n)$ be iid samples from probability mass function $p_X(x|\theta)$ (if X is discrete), or from density $f_X(x|\theta)$ (if X is continuous), where θ is a parameter (or vector of parameters).
- We define the likelihood of x given θ to be the probability of observing x if the true parameter is θ.
- If X is discrete:

$$L(x|\theta) = \Pi_{i=1}^{n} p_X(x_i|\theta)$$

- If X is continuous:

$$L(x|\theta) = \Pi_{i=1}^{n} f_X(x_i|\theta)$$

- We can always maximize a monotone function of the liklihood.
- Taking $Log_e(L(x|\theta))$ mitigates the maximization problems very often.

הטכניון

ביה"ס ללימודי המשך של הטכניון ע"ש עזריאלי

Maximum Likelihood - Exersize

- Give the likelihoods for each of the samples, and find which value of θ maximizes the likelihood.

1. Suppose $(x_1, x_2, x_3) = (1, 0, 1)$ are iid samples from $Ber(p)$ (recall p is the probability of a success).
2. Suppose $(x_1, x_2, x_3, x_4) = (3, 0, 2, 7)$ are iid samples from $Poi(\lambda)$ (recall λ is mean number of events in a unit of time).
3. Suppose $(x_1, x_2, x_3) = (3.22, 1.81, 2.47)$ are iid samples from $Exp(\theta)$ (recall θ is the historical average number of events in a unit of time).

Maximum Likelihood - Exersize

1. Consructing the Log likelihood:

$$L(x|p) = \Pi_{i=1}^{3} P_X(x_i|p) = P(1|p)P(0|p)P(1|p) = p^2(1-p)$$

- Taking the derivative by p and equating to zero:
  $3p^2 - 2p = 0 \Rightarrow \hat{p} = \frac{2}{3}$

2. Consructing the Log likelihood:

$$L(x|p) = \Pi_{i=1}^{4} P_X(x_i|p) = P(3|\lambda)P(0|\lambda)P(2|\lambda)P(7|\lambda)$$

$$= e^{-\lambda}\frac{\lambda^3}{3!}e^{-\lambda}\frac{\lambda^0}{0!}e^{-\lambda}\frac{\lambda^2}{2!}e^{-\lambda}\frac{\lambda^7}{7!} = Ce^{-4\lambda}\lambda^{12}$$

- Taking log, summing the powers and taking derivative:

$$-4 + 12\frac{1}{\lambda} = 0 \Rightarrow \hat{\lambda} = \frac{12}{4}$$

Maximum Likelihood - Exersize

3.
$$L(x|p) = \Pi_{i=1}^{3} f_X(x_i|p) = f(x_1|p)f(x_2|p)f(x_3|p)$$
$$= \theta e^{-3.22\theta}\theta e^{-1.81\theta}\theta e^{-2.47\theta} = \theta^3 e^{-7.5}$$

- Taking log and derivative we get:

$$\hat{\theta} = \frac{3}{7.5} = 0.4$$

Unbiasedness

- Let $\hat{\theta}$ be some estimator for θ.
- We say that $\hat{\theta}$ is unbiased if:

$$E(\hat{\theta}) = \theta$$

- Example: Is $\hat{\mu} = \bar{x}$ a unbiased estimator for μ?
- Answer: yes

$$E(\bar{x}) = \mu$$

Unbiasedness

- Is $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ unbiased for σ^2 where $\sigma^2 = Var(x)$.
- Answer: No

$$E(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2) = \frac{n-1}{n}\sigma^2$$

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

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

is unbiased.

Exercise

- Simulate a 10X1000 matrix with $x_1, x_2, \ldots, x_{10} \sim exp(2)$
- Calculate a biased and unbiased estimator for $Var(X) = \frac{1}{\lambda^2}$
- Average on the results for both methods.