1. Sample Statistic

A statistic is a function of the available data t=T(X_1,X_2,…,X_n). it is important because it summarize some aspects of the underlying data, and it can ba calculated even if the parameter is unknown: for the sample mean it is the sum of the observed x divided by their number, while the sample variance is the deviance of the observed x and the sample mean of the x over the number of the observations -1. The former gives information about the location of the observed distribution, and the latter about the scale. The problem is that the statistics change every time we change the observed result of the experiment, so we should apply the function T not on the observed value but on the r.v. and it is the SAMPLE STATISTIC.

We do not know the distribution of each rv so event the sample statistic is a rv and it has a probability distribution called SAMPLING DISTRIBUTION and using it to evaluate how far the sample statistic is form the population parameter. It allows us to make inference!

The sampling distribution depends on the population distribution, so we could derive it, for example the sample mean in a Bernoulli population is equal to the sum of the Bernoullian rv divided by their number, knowing that the sum of the bernoullian rv are distributed as a binomial.

Population distribution → distribution of the elements of the population;
Sample distribution → distribution of the element of the sample drawn from a population;
Sampling distribution → behavior of a statistic referring to a large number of rv samples.

2. Random Sample - Observed Sample

We know that a population is a collection of the objects under investigation and it is divided in two kind: finite such as the shareholders of the a company described with the frequency distribution and infinite such as the future buyers of that share described with a pdf; while a sample is a part of the population that is representative of all the population containing all the relevant information about that, described with the frequency distribution called sample distribution. The multiple rv (X_1,X_2,,…,X_n) is a RANDOM SAMPLE, and one of its possible realization is the OBSERVED SAMPLE.

3. Binomial - Bernoulli Distribution

The Bernoulli distribution is a dichotomous variable with result equals to 0 or 1, success or failure, it depends only on the value of the probability of the experiment that is the unknown parameter in this case; the binomial distribution can be seen as the sum of different Bernoulli trials, and it is distributed depending on two parameter the number of trials *n*, and the probability of each trial to be successfully theta.

4. Samplig Distribution of sample mean/variance

Applying the function T not on the observed value but on the r.v. and it is the SAMPLE STATISTIC.

We do not know the distribution of each rv so event the sample statistic is a rv and it has a probability distribution called SAMPLING DISTRIBUTION and using it to evaluate how far the sample statistic is form the population parameter. It allows us to make inference!

The sampling distribution depends on the population distribution, so we could derive it, for example the sample mean in a Bernoulli population is equal to the sum of the Bernoullian rv divided by their number, knowing that the sum of the bernoullian rv are distributed as a binomial.

Population distribution → distribution of the elements of the population;
Sample distribution → distribution of the element of the sample drawn from a epopulation;
Sampling distribution → behavior of a statistic referring to a large number of rv samples.

5. Likelihood function of poisson

Given the observed sample **x**=(x_1,x_2,…,x_n) generated by a statistical model M, we could wonder which is the probability of observing that particular result. Since the observed result is composed by n result of n different rv iid, we should do the intersection of each probability to obtain that particular result so the product of all them. The likelihood is not a probability (it could be in the discrete case) but a likelihood as we said. It is a function that goes from the parametric space CAPITAL THETA to infinite

$$f_\lambda(x) = \frac{e^{-\lambda}}{x!}\lambda^x$$

$$L(\lambda; x) = \prod_{i=1}^{n} \frac{e^{-\lambda}}{x_i!} \lambda^{x_i} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{m} x_i}}{\prod_{i=1}^{m} x_i!} \propto e^{-n\lambda} \lambda^{\sum_{i=1}^{m} x_i}$$

6. Maximum likelihood estimate – properties

If T is a ML estimator of theta:

T is invariant;
If S is a sufficient statistic for theta, the ML estimator of theta is a function of S;
If there exists an unbiased and efficient estimator of theta, it is the ML estimator;
T is asymptotically unbiased and efficient (its variance asymptotically attains the Cramer-Rao lower bound), hence it is constant;
T is asymptotically normally distributed (consistent and asymptotically distributed CAN).

7. Fisher information (score/observed function)

For the maximum likelihood estimate two conditions must hold: the first derivative is equal to zero in that point and the second derivative is <0 when computing in that point; we can also use the transformed likelihood function, the log-likelihood function because the logarithm is a monotonically increasing function and the point of maximum coincides in both the function.

Before the realization of the rv the first derivative of the log-likelihood function depends also on the **X** and it is called SCORE FUNCTION and it is the first derivative with respect to theta of the log-likelihood function. The expected value of the score function is 0 and the variance of the score function is the FISHER INFORMATION. The expected value of the second derivative of the log-likelihood function is equal to the fisher information.

The second derivative of the log-likelihood with the negative sign is the OBSERVED INFORMATION denoted by J. it is the Hessian matrix of the log-likelihood function and the Jacobian matric of the score function.

8. Sufficient Statistics

We know that a statistic summarizes important aspect but there is also a loss of information, two sample statistic can be equal even if the two samples are different, so we must find the statistic that allows us to see the statistic or the observed sample and have the same set of information. They are the SUFFICIENT STATISTICS. A statistic T(x) is called sufficient for theta if the conditional sampling distribution of X=(X_1,X_2,…, X_n) given T(X) does not depend on theta:

$$f(x_1, x_2, \dots, x_n | T(x) = t) = \frac{f_\theta(x_1, x_2, \dots, x_n, T(x))}{g_\theta(t)} = \frac{f_\theta(x_1, x_2, \dots, x_n)}{g_\theta(t)}$$

Must not depend on theta.

9. Estimator Properties

If a statistic is used to estimate a parameter, it is called ESTIMATOR, its realization is the pint estimate. The first property of an estimator is the that it is unbiasedness, an estimator T of the parameter that is unbiased if $E_\theta(T) = \theta$

If the estimator is biased if the difference $B(\theta) = E_\theta(T) - \theta$ is not null.

MSE (mean square error and efficiency) is

$$MSE_\theta(T) = E_\theta[(T - \theta)^2] = Var_\theta(T) + [B(\theta)]^2$$

So if an estimator is unbiased the MSE is equal to the variance of the T, the T* with the lowest variance is said to be the UNIFORMLY MINIMUM VARIANCE UNBIASED ESTIMATOR (UMVUE). To find it we can use the Cramer-Rao Inequality.

Increasing the sample size we will improve the knowledge about the behavior of the estimator; asymptotically unbiasedness; consistency; asymptotic normality.

10. Hypothesis testing (schema 4 situazioni)

A parametric statistical hypothesis is a statement about one or more values that a given parameter can assume, or relationships between parameters.

The hypothesis of interest is the null hypothesis H0, where theta belongs to theta0 and the other is the alternative hypothesis where H1 theta belongs to theta1, theta0 + theta1 are equal to all the parametric space theta usually, and the intersection between the two is an empty set. We can have hypothesis simple or composite. After the experiment and after defining the two hypothesis we can denote two kind of region in the sample space: the acceptance region and the critical one.

State Action	H0 true	H0 false
We accept H0	ok	II type error
We reject H0	I type error	ok

Alpha is probability of I type error; Beta is the probability of II type error.

Alpha is the significance level; pi-greek = 1- Beta is the power test. The power of the test is the probability of correctly rejecting the null.

Most powerful test if pi-greek* is equal to alpha level and so pi-greek is the greatest pi-greek between all the others.

11. Power Function (with graph)

State Action	H0 true	H0 false
We accept H0	ok	II type error
We reject H0	I type error	ok

Alpha is probability of I type error; Beta is the probability of II type error.

Alpha is the significance level; pi-greek = 1- Beta is the power test. The power of the test is the probability of correctly rejecting the null.

Most powerful test if pi-greek* is equal to alpha level and so pi-greek is the greatest pi-greek between all the others.

12. Neyman pearson lemma

H0: theta = theta0 vs H1: theta = theta1. → two simple hypothesis

If:

Rejection region is equal to the **x** such that the likelihood function under the alternative over the likelihood function under the null is > k; and the probability of X belonging to the rejection region under the null is equal to alpha. Then R is the most powerful size alpha critical region.

13. P-value

If we base are tests on different level of alpha we can end up with different decisions. To avoid this arbitrariness, we could consider a critical region with bound equal to the observed sample mean and then measure its size. So this implies, considering the probability that under the null the test statistic is at least as large as its observed

value, the p-value. Smaller p-value, the more unusual the data would be under the null.

14. OLS estimator

In order to estimate the parameter of the regression we can use the ordinary least square method, it is a function of beta0 and beta1, so a sum of the (yi-beta0-beta1xi)^2, if we denote the betas with the hat then we can end up to the sum of the e_i^2 that we want to minimize. So we produce the system of the partial derivative with respect to beta0 and beta1 hat (normal equations).

The only condition that we can derive from the second normal equation is that the sample deviance of the x should not be equal to 0, so at least two observed value of the x should be different, otherwise all the values of the sample are equal!

15. Proprietà algebrica degli stimatori (slide 57)

The system of normal equations admits always a solution and it is unique; the first normal equation shows that the sum of the residual is null (and if we divide the sum of the residual by n and so we did the same for the (yi-yihat) we can see that the mean of the observed is equal to the mean of the estimated); the barycenter belongs to the distribution from the formula of beta0 hat; from the second normal equation we can see that the mean of X * e is equal to the covariance between X and e and it is equal to 0.

16. Decomposition of the deviance (con passaggi)

We need to establish how the model chosen with the ols criterion is able to fit. So we can study the deviance of Y and decompose it:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSE + SSR$$

(vedere dimostrazione completa)

SSE = sum of squared errors, deviance of the residuals
SSR = sum of squares of regressions and represents the part of the total variation of Y that is explained by the regression.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

In the simple linear regression model $R^2 = r_{xy}^2$

In the multiple linear regression model $R^2 = r_{y\hat{y}}^2$

17. OLS estimator of beta

We follow the same procedure as before trying to minimize the squared of the difference between the observed value of y and the estimated, in order to have the estimation of beta0 and beta1, so also in this case we should minimize the square of the sum of the estimated residuals and do the derivative, we recall that this time we have a column vector, so to have the square we do:

e'e = (y-Xbeta)'(y-Xbeta) = y'y – 2y'Xbeta + beta'X'Xbeta

we do the first derivative with respect to beta to obtain:

beta hat = (X'X)^-1 X'y providing the full rank for r(X)=r(X'X)=k

and the second derivative is equal only to X'X that prove that it is positive so the beta hat are point of minimum.

18. Gauss – Markov Theorem

Under the assumption:

I) Y=Xbeta + epsilon
II) E(epsilon|X) = 0
III) r(X) = r(X'X) = k
IV) Var(epsilon|X) = sigma^2 I_n

Among all the possible linear and unbiased estimators of beta, OLS estimator is the best so it is called BLUE.

19. Statistic properties of estimator

B is the estimator of beta and b its realization in the observed sample.

B is a linear estimator, in fact, we can see that

$$B = (X'X)^{-1}X'Y = QY$$

The expected value is:

$$E(B|X) = E[QY|X) = \beta + QE[\varepsilon|X] = \beta$$

The estimator is unbiased, independently of its distribution and of the sample size.

$$B = \beta + (X'X)^{-1}\varepsilon\,,$$

$$Var(B|X) = E[(B-\beta)(B-\beta)'|X] = E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X]$$

$$Var(B|X) = (X'X)^{-1}\sigma^2$$

We have to estimate the variance sigma:

$$\hat{\sigma}^2 = S^2 = \frac{e'e}{n-k} = \frac{SSE}{n-k}\ it\ is\ an\ unbiased\ estimator!!$$

20. t-test and f-test (come si distribuiscono)

t-test

$$Z = \frac{B_j - \beta_j}{sd(B_j)} \sim N(0,1)$$

$$sd(B_j) = \sqrt{Var(B_j)} = \sigma\sqrt{[(X'X)^{-1}]_{jj}} = \frac{\sigma}{\sqrt{dev(X_j)(1-R_j^2)}}$$

where the R^2_j is the coefficient of determination in the regression of X_j on all the other independent variables. The complementary with 1 is the part of X_j not linearly explained by the other covarites.

Since the variance sigma is unknown, we estimate it through S^2 and it results:

$$T = \frac{B_j - \beta_j}{se(B_j)} \sim t_{n-k}\ and\ se(B_j) = \sqrt{\widehat{Var(B_j)}} = \frac{s}{\sqrt{dev(X_j)(1-R_j^2)}}$$

The confidence interval will be:

$$\Pr\left[-t_{n-k;\frac{\alpha}{2}} \le \frac{B_j - \beta_j}{se(B_j)} \le t_{n-k;\frac{\alpha}{2}}\right] = 1 - \alpha$$

f-test

if we need to test multiple hypothesis about the parameters? We know that the most parsimonious model will have a residual deviance greater than the unrestricted. The question is whether the increase in the SSE is statistically significant or not.

For q exclusion:

$$\frac{SSE_r - SSE_u}{q} \sim \sigma^2 \chi_q^2$$

Where q = df_r – df_u = n-(k-q) – (n-k)

$$S^2 = \frac{SSE_u}{n-k} \sim \frac{\sigma^2}{n-k} \chi_{n-k}^2$$

$$F = \frac{\frac{SSE_r - SSE_u}{q}}{\frac{SSE_u}{n-k}} \sim F_{q;n-k}$$

F_q is the Fisher distribution and we know that SSE=SST(1-R^2)

$$F = \frac{\frac{R_u^2 - R_r^2}{q}}{\frac{1 - R_u^2}{n-k}}$$

q=1 → F=T^2

q=n-1 → ANOVA test $F = \frac{R^2(n-k)}{q(1-R^2)}$

21. Proxy and instrumental variable

In the model mis-specification in particular for the underfitting problem.

The omission of relevant variable could lead to the biasedness and inconsistency of the estimator B, because the disturbance term now must include the omitted variable, and the estimator B_1 would remain unbiased only if the co-deviance between the explanatory variable is 0, so if they are uncorrelated. So to solve this problem we should check the restrictions made on the model and evaluate the parameter through the use of an F-test. If the variables entering in the model re unknown? We could

introduce resorting to proxies, but it can reduce the power of the test , an alternative solution is that using the instrumental variables and IV estimator

22. VIF → Varianza dello stimatore di β

In the overfitting specification, the estimates are always unbiass and consistent but the variability increases

$$Var(B_j) = \frac{\sigma^2}{dev(X_j)(1 - R_j^2)}$$

If the X_j is irrelevant its informative content is already present in the model in the other regressors, so it would be linearly correlated with them. The higher the correlation, the higher the R^2, the higher the denominator and so the variance of B_j

$$Variance\ Inflation\ Factor = VIF_j = \frac{1}{1 - R_j^2}$$

If VIF_j=1 they are uncorrelated otherwise it takes higher value (the maximum threshold is 2 so when R^2 is 0.5)

23. Chow Test

We are in the mis-specification of the structural instability. In order to check the hypothesis of the consistency of the functional relationship, the CHOW test is a particular F-test, dividing the sample in two parts: so with two different SSE but with the same variance for the two part and the it can be compared with the full model without restriction:

$$F = \frac{\frac{SSE_r - SSE_u}{k}}{\frac{SSE_u}{n_I + n_{II} - 2k}} \sim F_{k;n-2k}$$

The two sample variance both constitute an unbiased estimator of sigma^2

$$F = \frac{\frac{s_I^2}{\sigma^2}}{\frac{s_{II}^2}{\sigma^2}} \sim F_{n_I - k, n_{II} - k}$$

24. Durbin – Watson

The presence of autocorrelation implies a lose in the reliability of the inferential procedure. The most used procedure is the DW test:

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{i=1}^{n} e_t^2} \sim 2 - 2\hat{\rho}$$

But it leaves a region of uncertainty, so it is not the most efficient.

25. GLS

Generalized least squares estimators we know that the matrix of the variance is multipled by the matrix V because now it contains values different from 0 also outside the main diagonal. We should know the matrix V. also with the feasible GLS estimator basing all the procedure on the shape of the matrix V, so V=V(X;theta), then it is possible to estimate the vector beta. The correct specification of the variance and covariance model of the disturbances and the strict exogeneity of the predictors are two fundamental assumptions for the consistency of the FGLS estimators. To overcome this limit it is possible to use the HAC (heteroscedasticity and autoregressive consistent), valid only for large sample.

26. Endogeneità e esogeneità

We saw that to evaluate the effect of a given regressor, by holding all the other potentially influential factors constant, the random component epsilon should be independent of it. It is necessary to estimate beta1. The imposition of the Cov(X,epsilon)=0 is not sufficient because there could be other non linear linkage. A more strong condition is that E(epsilon|X)=E(epsilon)=0, which implies the orthogonality condition E(epsilon*X)=0. This means that the regressor has to be exogenous!!!

27. CAPM

The equilibrium return for risky assets in condition of perfect market (rational, risk adverse, perfectly informed, mean-variance criterion and no trading costs).
Risk for the asset *specific;* risk of the market or *systematic* not diversifiable.

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f]$$

Or

$$E(R_i) - R_f = \beta_i[E(R_m) - R_f]$$

β_i is the sensitivity of the expceted asset returns to the expected excess market returns

$$\beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)} = \frac{\rho_{i,m}\,\sigma_i}{\sigma_m}$$

Beta>1 more risky securities and higher returns
Beta<1 lower returns
Beta = 0 returns independent from the market

$$E(R_i) = R_f + \beta_i\left[E(R_m) - R_f\right]$$
$$E(R_i) = \left[R_f - \beta_i R_f\right] + \beta_i R_f$$