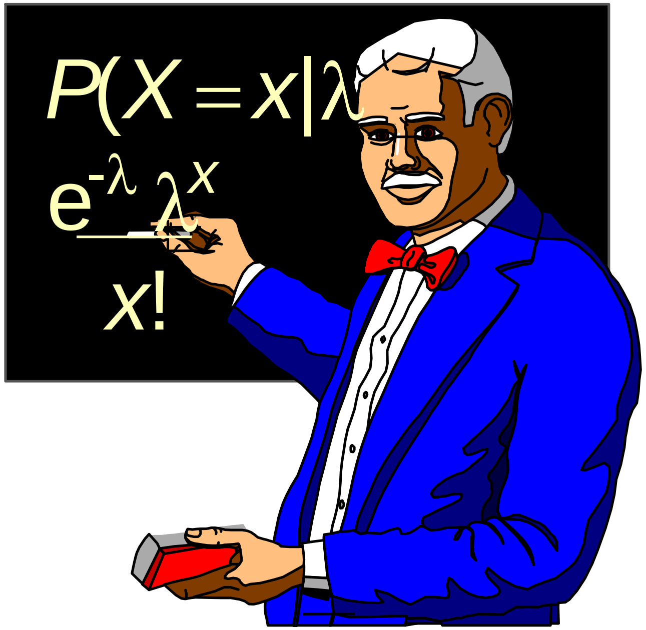
Phys 443
Computational Physics

Regression Analysis



Chi-square test

The **Chi-square test statistic** is:

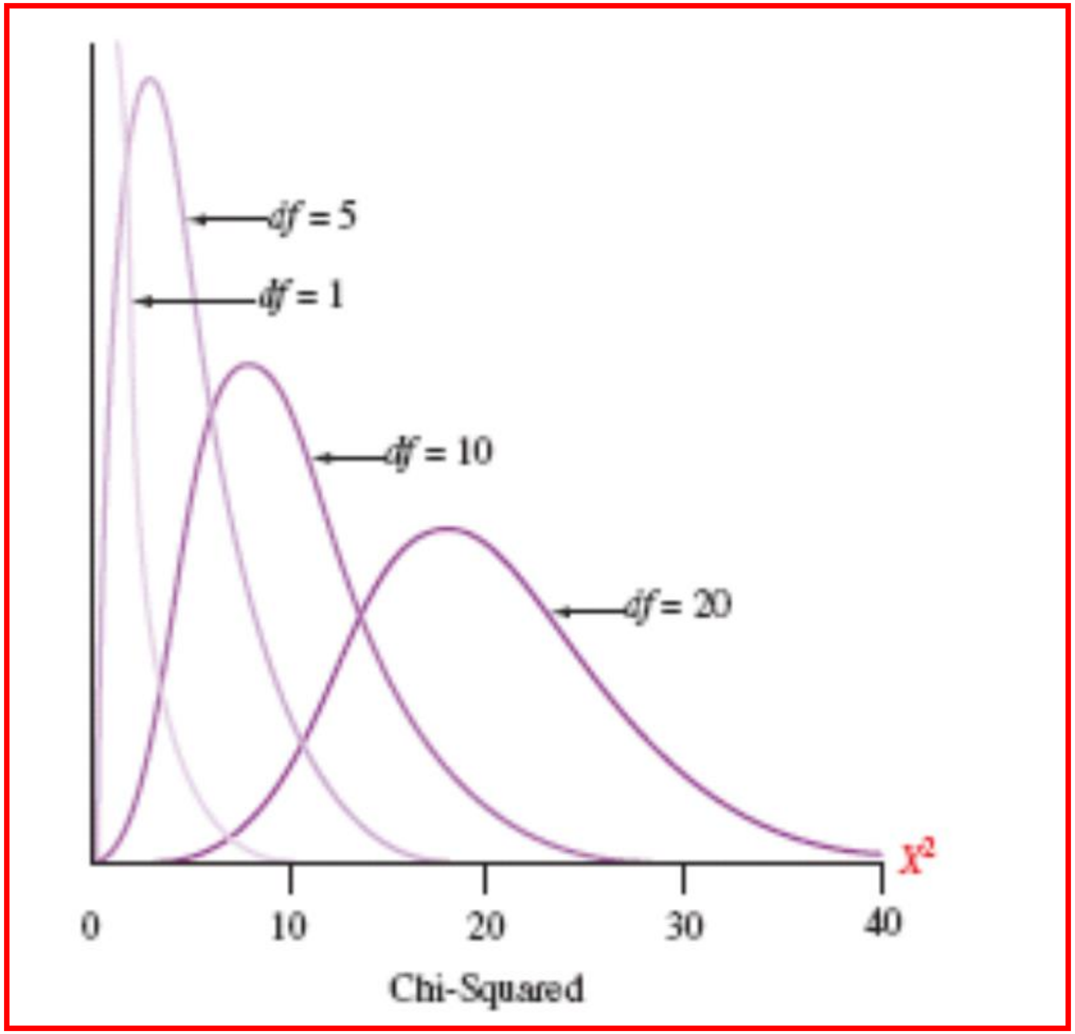
$$\chi^2 = \sum_{all\ \text{cells}} \frac{(Obs - Exp)^2}{Exp}$$

where:
O*bs* = observed frequency
Exp= expected frequency if H_0 is true

The expected frequency *i* is np_i

- **Large values of χ^2 are evidence against H_0 because they say the observed counts are far from what we would expect if H_0 were true.**
- **Chi-Square tests are one-side (even though H_A is many-sided)**

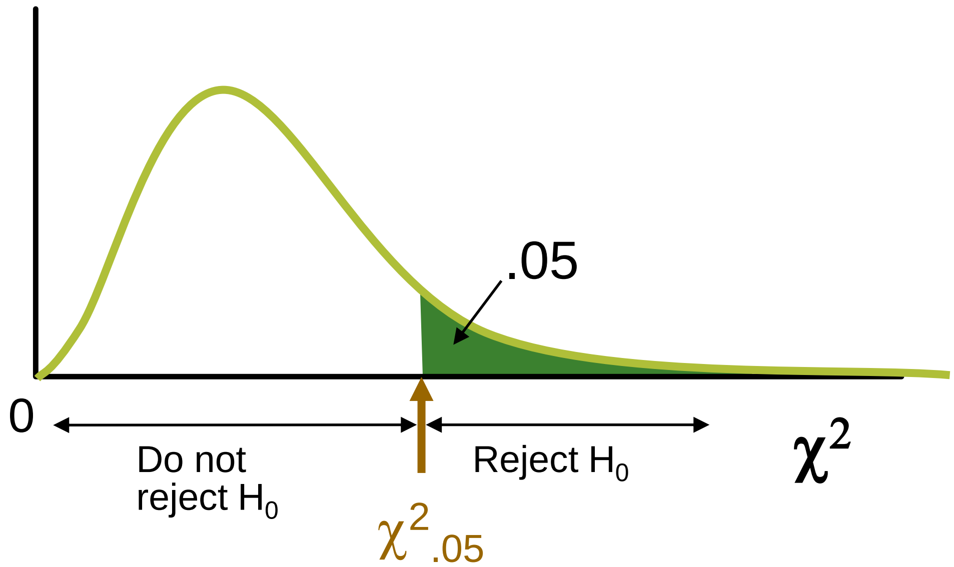
Chi-square test



Chi-square test

- The χ^2 test statistic approximately follows a chi-squared distribution with *k*-1 degrees of freedom, where *k* is the number of categories.

Decision Rule:
If $\chi^2 > \chi^2_{.05}$, reject H_0,
otherwise, do not reject H_0.



Chi-square test

ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,4549	1,0742	1,6424	2,7055	3,8415	5,4119	6,6349	10,8274
2	0,2107	1,3863	2,4079	3,2189	4,6052	5,9915	7,8241	9,2104	13,8150
3	0,5844	2,3660	3,6649	4,6416	6,2514	7,8147	9,8374	11,3449	16,2660
4	1,0636	3,3567	4,8784	5,9886	7,7794	9,4877	11,6678	13,2767	18,4662
5	1,6103	4,3515	6,0644	7,2893	9,2363	11,0705	13,3882	15,0863	20,5147
6	2,2041	5,3481	7,2311	8,5581	10,6446	12,5916	15,0332	16,8119	22,4575
7	2,8331	6,3458	8,3834	9,8032	12,0170	14,0671	16,6224	18,4753	24,3213
8	3,4895	7,3441	9,5245	11,0301	13,3616	15,5073	18,1682	20,0902	26,1239
9	4,1682	8,3428	10,6564	12,2421	14,6837	16,9190	19,6790	21,6660	27,8767
10	4,8652	9,3418	11,7807	13,4420	15,9872	18,3070	21,1608	23,2093	29,5879
11	5,5778	10,3410	12,8987	14,6314	17,2750	19,6752	22,6179	24,7250	31,2635
12	6,3038	11,3403	14,0111	15,8120	18,5493	21,0261	24,0539	26,2170	32,9092
13	7,0415	12,3398	15,1187	16,9848	19,8119	22,3620	25,4715	27,6882	34,5274
14	7,7895	13,3393	16,2221	18,1508	21,0641	23,6848	26,8727	29,1412	36,1239
15	8,5468	14,3389	17,3217	19,3107	22,3071	24,9958	28,2595	30,5780	37,6978
16	9,3122	15,3385	18,4179	20,4651	23,5418	26,2962	29,6332	31,9999	39,2518
17	10,0852	16,3382	19,5110	21,6146	24,7690	27,5871	30,9950	33,4087	40,7911
18	10,8649	17,3379	20,6014	22,7595	25,9894	28,8693	32,3462	34,8052	42,3119
19	11,6509	18,3376	21,6891	23,9004	27,2036	30,1435	33,6874	36,1908	43,8194
20	12,4426	19,3374	22,7745	25,0375	28,4120	31,4104	35,0196	37,5663	45,3142
21	13,2396	20,3372	23,8578	26,1711	29,6151	32,6706	36,3434	38,9322	46,7963
22	14,0415	21,3370	24,9390	27,3015	30,8133	33,9245	37,6595	40,2894	48,2676
23	14,8480	22,3369	26,0184	28,4288	32,0069	35,1725	38,9683	41,6383	49,7276
24	15,6587	23,3367	27,0960	29,5533	33,1962	36,4150	40,2703	42,9798	51,1790
25	16,4734	24,3366	28,1719	30,6752	34,3816	37,6525	41,5660	44,3140	52,6187
26	17,2919	25,3365	29,2463	31,7946	35,5632	38,8851	42,8558	45,6416	54,0511
27	18,1139	26,3363	30,3193	32,9117	36,7412	40,1133	44,1399	46,9628	55,4751
28	18,9392	27,3362	31,3909	34,0266	37,9159	41,3372	45,4188	48,2782	56,8918
29	19,7677	28,3361	32,4612	35,1394	39,0875	42,5569	46,6926	49,5878	58,3006
30	20,5992	29,3360	33,5302	36,2502	40,2560	43,7730	47,9618	50,8922	59,7022

Chi-square test for homogeneity

- Setting: We have several data sets (for example results of applying several different treatments.)
- Homogeneity (the null hypothesis) means that the data sets are all drawn from the same distribution: <u>that all the treatments are equally effective.</u>
- Three treatments for a covid-19 are compared in a clinical trial, yielding the following data:

	Treatment 1	Treatment 2	Treatment 3
Cured	50	30	12
Not cured	100	80	18

Use a chi-square test to compare the cure rates for the three treatments, i.e., to test if all three cure rates are the same.

Chi-square test for homogeneity

- H_0 = all three treatments have the same cure rate.
- H_A = the three treatments have different cure rates.
- Expected counts:

Under H_0 the cure rate is
(total cured)/(total treated) = 92/290 = 0.317

  - This gives the following table of observed and expected counts (observed in black, expected in blue).
  - We include the marginal values (in red). These were used to compute the expected counts.

	Treatment 1	Treatment 2	Treatment 3	
Cured	50, 47.6	30, 34.9	12, 9.5	92
Not cured	100, 102.4	80, 75.1	18, 20.5	198
	150	110	30	290

Chi-square test for homogeneity

Likelihood ratio statistic: $G = 2 \sum O_i \ln(O_i / E_i) = 2.12$

Pearson's chi-square statistic: $X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 2.13$

Degrees of freedom means how many choices describe the data.

Formula: degrees of freedom df = (2 -1) (3 - 1) = 2.

p-value = 0.346, $\alpha = 0.005$

$p > \alpha$

The data does not support rejecting H_0. We do not conclude that the treatments have different efficiency

Chi-square test for homogeneity

XX is thinking of buying a restaurant and asks about the distribution of lunch customers. The owner provides row one below. XX records the data in row two himself one week.

	M	T	W	R	F	S
Owner's distribution	.1	.1	.15	.2	.3	.15
Observed # of cust.	30	14	34	45	57	20

Run a chi-square goodness-of-fit test on the null hypotheses:
H_0: the owner's distribution is correct.
H_A: the owner's distribution is not correct.
Compute X^2.

Chi-square test for homogeneity

The total number of observed customers is 200.
The expected counts (under H_0) are 20 20 30 40 60 30

$$X^2 = \sum \frac{(O_i - E_i)^2|}{E_i} = 11.44$$

df = 6 - 1 = 5 (6 cells, compute 1 value -the total count- from the data)

p = 0.043.

So, at a significance level of 0.05 we reject the null hypothesis in favor of the alternative that the owner's distribution is wrong.

Chi-square test for homogeneity

Table 2 (cont'd). One-sided P-values from $\chi^2(\nu)$ distribution: $P[\chi^2(\nu) > c]$.

	$df = \nu$									
c	1	2	3	4	5	6	7	8	9	10
8.5	0.004	0.014	0.037	0.075	0.131	0.204	0.291	0.386	0.485	0.580
8.6	0.003	0.014	0.035	0.072	0.126	0.197	0.283	0.377	0.475	0.570
8.7	0.003	0.013	0.034	0.069	0.122	0.191	0.275	0.368	0.465	0.561
8.8	0.003	0.012	0.032	0.066	0.117	0.185	0.267	0.359	0.456	0.551
8.9	0.003	0.012	0.031	0.064	0.113	0.179	0.260	0.351	0.447	0.542
9.0	0.003	0.011	0.029	0.061	0.109	0.174	0.253	0.342	0.437	0.532
9.2	0.002	0.010	0.027	0.056	0.101	0.163	0.239	0.326	0.419	0.513
9.4	0.002	0.009	0.024	0.052	0.094	0.152	0.225	0.310	0.401	0.495
9.6	0.002	0.008	0.022	0.048	0.087	0.143	0.212	0.294	0.384	0.476
9.8	0.002	0.007	0.020	0.044	0.081	0.133	0.200	0.279	0.367	0.458
10.0	0.002	0.007	0.019	0.040	0.075	0.125	0.189	0.265	0.350	0.440
10.2	0.001	0.006	0.017	0.037	0.070	0.116	0.178	0.251	0.335	0.423
10.4	0.001	0.006	0.015	0.034	0.065	0.109	0.167	0.238	0.319	0.406
10.6	0.001	0.005	0.014	0.031	0.060	0.102	0.157	0.225	0.304	0.390
10.8	0.001	0.005	0.013	0.029	0.055	0.095	0.148	0.213	0.290	0.373
11.0	<.001	0.004	0.012	0.027	0.051	0.088	0.139	0.202	0.276	0.358
11.2	<.001	0.004	0.011	0.024	0.048	0.082	0.130	0.191	0.262	0.342
11.4	<.001	0.003	0.010	0.022	0.044	0.077	0.122	0.180	0.249	0.327
11.6	<.001	0.003	0.009	0.021	0.041	0.072	0.115	0.170	0.237	0.313
11.8	<.001	0.003	0.008	0.019	0.038	0.067	0.107	0.160	0.225	0.299

The F Test: example

Consider the following table of counts

Use a chi-square test with significance level 0.01 to test the hypothesis that the number of marriages and education level are independent.

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

The F Test: example

The null hypothesis is that the cell probabilities are the product of the marginal probabilities. Assuming the null hypothesis we estimate the marginal probabilities in red and multiply them to get the cell probabilities in blue.

Education	Married once	Married multiple times	Total
College	0.365	0.061	611/1436
No college	0.492	0.082	825/1436
Total	1231/1436	205/1436	1

We then get expected counts by multiplying the cell probabilities by the total number of women surveyed (1436). The table shows the observed, expected counts:

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

The F Test: example

We then have
$G = 16.55$ and $X^2 = 16.01$

The number of degrees of freedom is $(2 - 1)(2 - 1) = 1$. We could count this: we needed the marginal probabilities to compute the expected counts. Now setting any one of the cell counts determines all the rest because they need to be consistent with the marginal probabilities. We get $p = 0.000047$

Therefore we reject the null hypothesis in favor of the alternate hypothesis that number of marriages and education level are not independent

Regression Analysis

Basic idea:

- Use data to identify relationships among variables and use these relationships to make predictions.

Data

Statistics

Data

Information

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

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

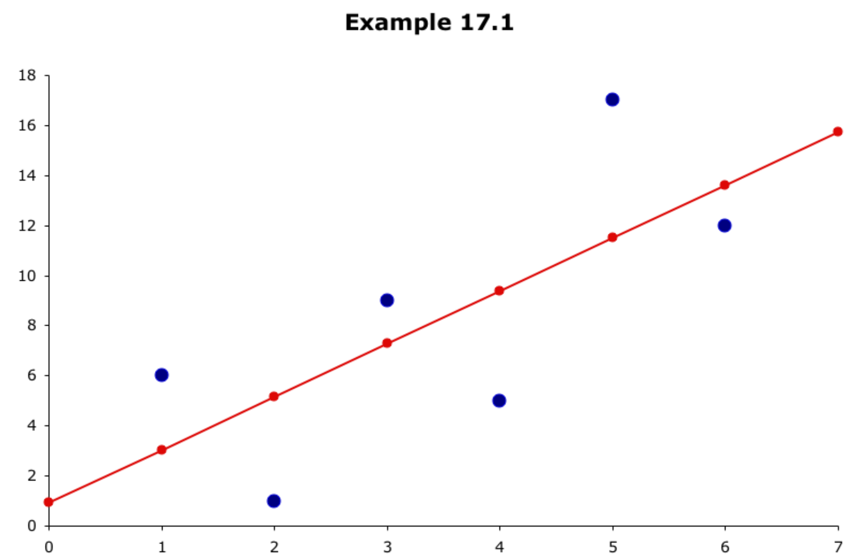
$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

Data Points:

x	y
1	6
2	1
3	9
4	5
5	17
6	12



$\hat{y} = .934 + 2.114x$

Linear regression

- Linear dependence: constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).
- **Regression analysis describes the relationship between two (or more) variables.**
- Example
  - For a conductor : Voltage versus Current

    Velocity versus time

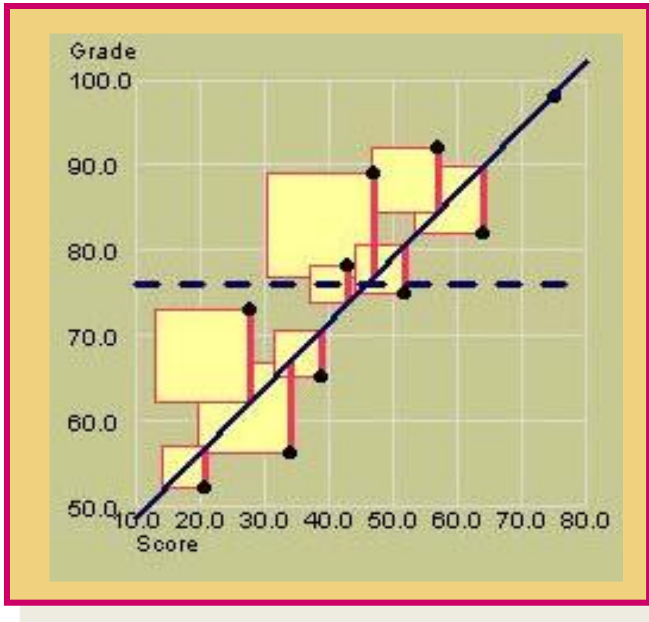
Steps in Regression Analysis

When you perform simple regression analysis, use a step-by step approach:

1. Fit the model to data – estimate parameters.
2. Determine how well the model fits the data.
3. Proceed to estimate or predict the quantity of interest

The Method of Least Squares

- The equation of the best-fitting line is calculated using *n* pairs of data (x_i, y_i).
- We choose our estimates $\hat{\alpha}$ and $\hat{\beta}$ to estimate α and β so that the vertical distances of the points from the line, are minimized.



$$\text{Best fitting line}: \hat{y} = \hat{\alpha} + \hat{\beta}x$$

$$\text{Choose } \hat{\alpha} \text{ and } \hat{\beta} \text{ to minimize}$$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Sum of Squares of Error(SSE)

Least Squares Estimators

Compute $\bar{x} = \dfrac{\sum x_i}{n}$, $\bar{y} = \dfrac{\sum y_i}{n}$,

$$S_{xx} = \sum x_i^{\ 2} - \frac{(\sum x_i)^2}{n}, \quad S_{yy} = \sum y_i^{\ 2} - \frac{(\sum y_i)^2}{n},$$

$$S_{xy} = \sum x_i \ y_i - \frac{(\sum x_i)(\sum y_i)}{n}.$$ Then

$$\hat{\beta} = \text{point estimator of } \beta = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \text{point estimator of } \alpha = \bar{y} - \hat{\beta}\ \bar{x}$$

Example: Age and Fatness

The following data was collected in a study of age and fatness in humans.

Age	23	23	27	27	39	41	45	49	50
% Fat	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1

Age	53	53	54	56	57	58	58	60	61
% Fat	34.7	42	29.1	32.5	30.3	33	33.8	41.1	34.5

One of the questions was, "What is the relationship between age and fatness?"

Example: Age and Fatness

Age (x)	% Fat y	x^2	xy
23	9.5	529	218.5
23	27.9	529	641.7
27	7.8	729	210.6
27	17.8	729	480.6
39	31.4	1521	1224.6
41	25.9	1681	1061.9
45	27.4	2025	1233
49	25.2	2401	1234.8
50	31.1	2500	1555
53	34.7	2809	1839.1
53	42	2809	2226
54	29.1	2916	1571.4
56	32.5	3136	1820
57	30.3	3249	1727.1
58	33	3364	1914
58	33.8	3364	1960.4
60	41.1	3600	2466
61	34.5	3721	2104.5
834	515	41612	25489.2

$$n = 18$$

$$\sum X = 834$$

$$\sum y = 515$$

$$\sum X^2 = 41612$$

$$\sum XY = 25489.2$$

Example: Age and Fatness

$$n = 18, \sum x = 834, \sum y = 515$$

$$\sum x^2 = 41612, \sum xy = 25489.2$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$= 41612 - \frac{834^2}{18} = 2970$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

$$= 25489.2 - \frac{(834)(515)}{18} = 1627.53$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$= \frac{1627.53}{2970}$$

$$= .55$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$= \frac{515}{18} - .55\frac{834}{18}$$

$$= 3.22$$

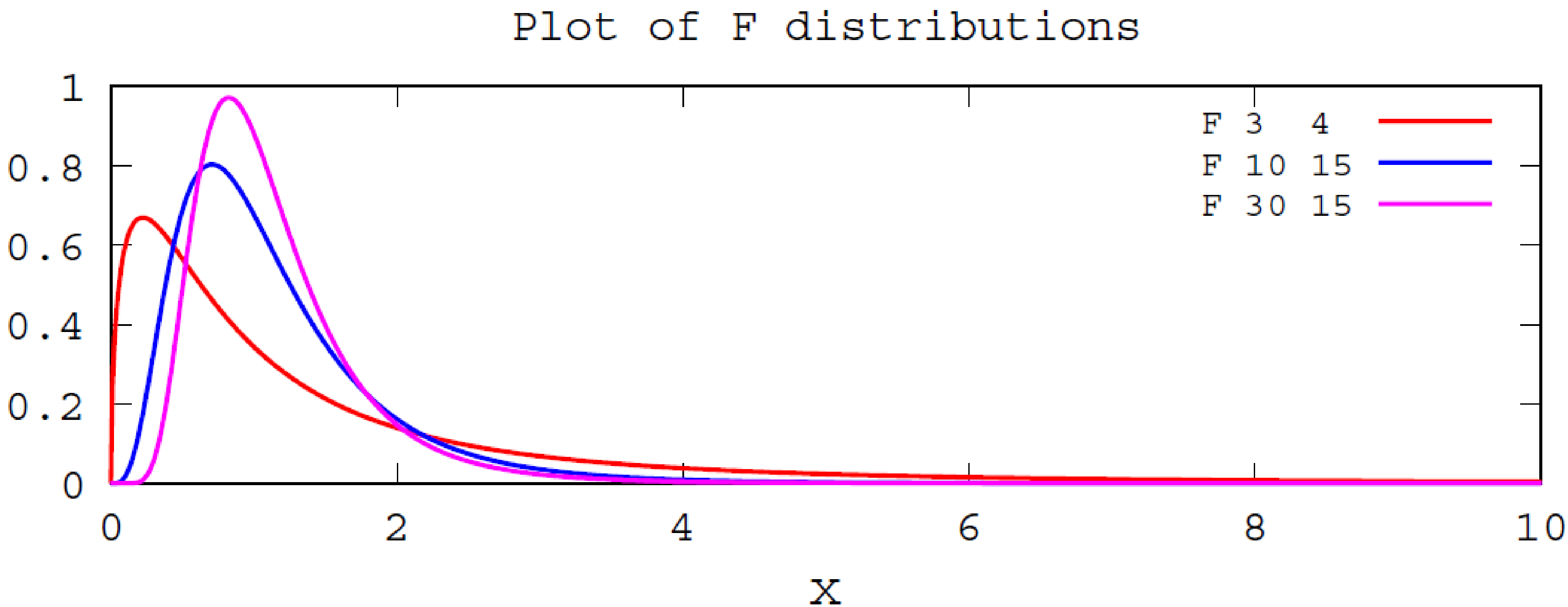
$$\hat{y} = 3.22 + .55x$$

F-Test

Notation: $F_{a,b}$, a and b degrees of freedom

Derived from normal data

Range: $[0,\infty)$



F-test = one-way ANOVA

Table 4

FIGURE 17.2 Sampling Distribution of F When H_0 Is True

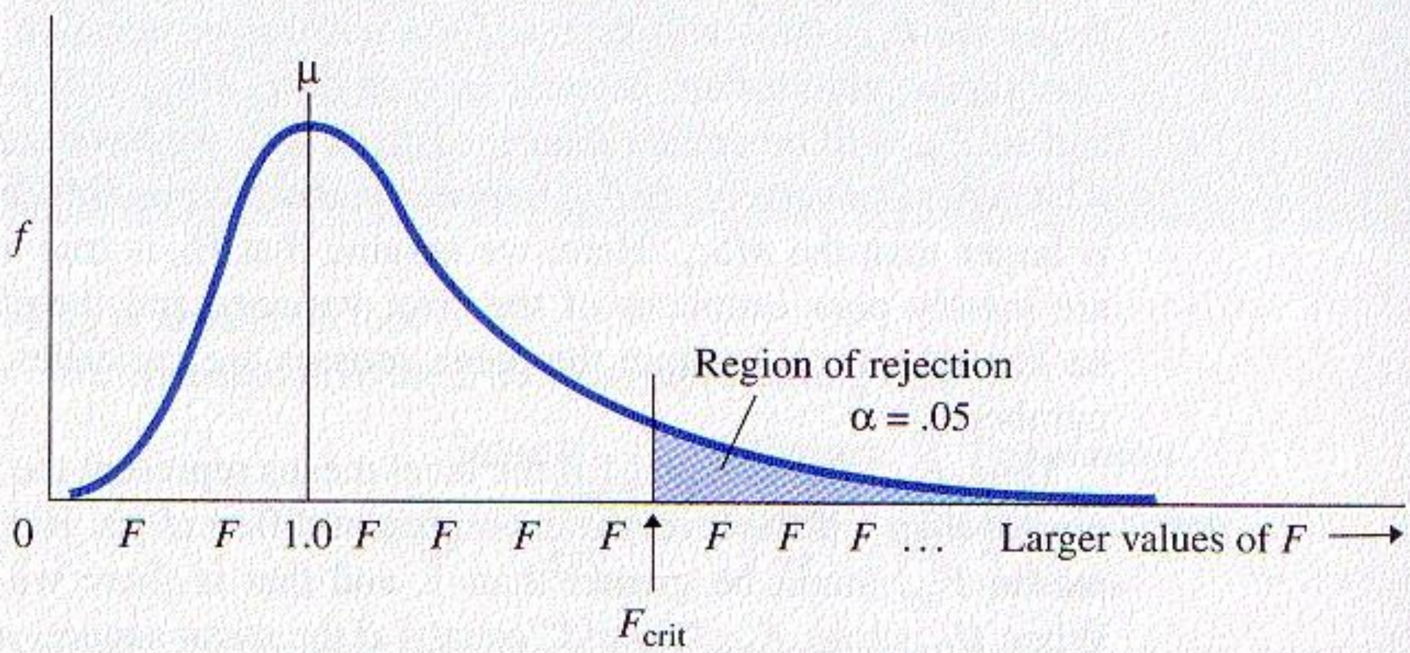


Table 4

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

The Analysis of Variance

- The total variation in the experiment is measured by the **total sum of squares**:

$$\text{Total SS} = S_{yy} = \sum (y - \bar{y})^2$$

- The Total SS is divided into two parts:

✓SSR (sum of squares for regression): measures the variation explained by including the independent variable *x* in the model.

✓SSE (sum of squares for error): measures the leftover variation not explained by *x*.

$$SSR = \frac{(S_{xy})^2}{S_{xx}}$$

$$SSE = \text{Total SS} - SSR$$

The ANOVA Table

Total *df* = n -1

Regression *df* = 1

Error *df* = n –1 – 1 = n - 2

Mean Squares

MSR = SSR/1

MSE = SSE/(n-2)

Source	df	SS	MS	F
Regression	1	SSR	MSR	MSR/MSE
Error	n - 2	SSE	MSE	
Total	n - 1	Total SS		

The F Test

We can test the overall usefulness of the linear model using an F test. If the model is useful, MSR will be large compared to the unexplained variation, MSE.

$$H_0 : \beta = 0 \quad vs \quad H_a : \beta \neq 0$$

$$\text{Test Statistic} : F = \frac{MSR}{MSE}$$

$$\text{Reject } H_0 \text{ if } F > F_\alpha \text{ with } 1 \text{ and } n-2 \text{ df.}$$

The F Test: example

The table shows recovery time in days for three medical treatments.
1. Set up and run an F-test testing if the average recovery time is the same for all three treatments.
2. Based on the test, what might you conclude about the treatments?

T_1	T_2	T_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

For $\alpha = 0.05$, the critical value of $F_{2,15}$ is 3.68.

The F Test: example

H_0 is that the means of the 3 treatments are the same. H_A is that they are not.

Our test statistic w is computed following the procedure from a previous slide. We get that the test statistic w is approximately 9.25. The p-value is approximately 0.0024. We reject H_0 in favor of the hypothesis that the means of three treatments are not the same.

Coefficient of Determination

The coefficient of determination is defined as

$$r^2 = \frac{SSR}{\text{Total SS}} = 1 - \frac{SSE}{\text{Total SS}}$$

- r^2 is the square of correlation coefficient
- r^2 is a number between zero and one and a value close to zero suggests a poor model.
- It gives the proportion of variation in y that can be attributed to an approximate linear relationship between x and y.
- A very high value of r^2 can arise even though the relationship between the two variables is non-linear. The fit of a model should never simply be judged from the r^2 value alone.

Estimate of σ

An estimator of the variance σ^2 is

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n - 2} = \mathrm{MSE}$$

Thus, an estimator of the standard deviation σ is

$$\hat{\sigma} = \sqrt{\frac{\mathrm{SSE}}{n - 2}} = \sqrt{\mathrm{MSE}}$$

Example: Age and Fatness

$$\text{Total SS} = \sum y^2 - \frac{(\sum y)^2}{n} = 16156.3 - \frac{515^2}{18} = 1421.58$$

$$\text{SSR} = \frac{(S_{xy})^2}{S_{xx}} = \frac{1627.53^2}{2970} = 891.27$$

$$\text{SSE} = \text{Total SS - SSR} = 1421.58 - 891.27 = 529.71$$

$$r^2 = 1 - \frac{\text{SSE}}{\text{Total SS}} = 1 - \frac{529.71}{1421.58} = .627$$

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2} = \frac{529.71}{18-2} = 33.11$$

$$\hat{\sigma} = \sqrt{33.11} = 5.75$$

Example: Age and Fatness

An analysis of variance (ANOVA) Table

Source	df	SS	MS	F
Regression	1	891.27	891.27	26.94
Error	16	529.71	33.11	
Total	17	1421.58		

- With r^2=0.627 or 62.7%, we can say that 62.7% of the observed variation in %Fat can be explained by your regression model with human age.
- The magnitude of a typical sample deviation from the least squares line is about 5.75(%) which is reasonably large compared to the y values themselves.
- This would suggest that the model is only useful in the sense of provide a rough estimates for %Fat for humans based on age.

Inference Concerning the Slope β

- Do the data present sufficient evidence to indicate that y increases (or decreases) linearly as x increases?
- Is the independent variable *x* useful in predicting *y*?
- A no answer to above questions means that *y* does not change, regardless of the value of *x*. This implies that the slope of the line, β, is zero.

$$H_0: \beta = 0 \quad \text{versus} \quad H_a: \beta \neq 0$$

Sampling Distribution

When the four basic assumptions of the simple linear regression model are satisfied, the following are true:

1. The mean value of $\hat{\beta}$ is β. That is, $\hat{\beta}$ is unbiased
2. The standard deviation of the statistic $\hat{\beta}$ is $\frac{\sigma}{\sqrt{S_{xx}}}$
3. $\hat{\beta}$ has a normal distribution (a consequence of the error e being normally distributed)
4. The probability distribution of the standardized variable

$$\boxed{t = \frac{\hat{\beta}}{\hat{\sigma}/\sqrt{S_{xx}}}}$$

has the t distribution with df=n-2

Confidence Interval for β

When then four basic assumptions of the simple linear regression model are satisfied, a (1-α)100% confidence interval for β is

$$\boxed{\hat{\beta} \pm t_{\alpha/2}\, \hat{\sigma} / \sqrt{S_{xx}}}$$

where the *t* critical value is based on df = n - 2.

Example: Age and Fatness

A 95% confidence interval for β is

$$\hat{\beta} \pm t_{\alpha/2}\, \hat{\sigma} / \sqrt{S_{xx}} = .55 \pm 2.12 \times 5.75 / \sqrt{2970} = .55 \pm .22$$

or $(.33, .77)$

<u>Based on sample data, the %Fat increases .55% on average with one year of age, and we are 95% confident that the true increase per year is between 0.33% and 0.77%.</u>

Hypothesis Tests Concerning β

Step 1: Specify the null and alternative hypothesis

- $\mathbf{H_0: \beta = \beta_0}$ **versus** $\mathbf{H_a: \beta \neq \beta_0}$ (two-sided test)
- $\mathbf{H_0: \beta = \beta_0}$ **versus** $\mathbf{H_a: \beta > \beta_0}$ (one-sided test)
- $\mathbf{H_0: \beta = \beta_0}$ **versus** $\mathbf{H_a: \beta < \beta_0}$ (one-sided test)

Step 2: Test statistic

$$\boxed{t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} / \sqrt{S_{xx}}}}$$

Step 3: When four basic assumptions of the simple linear regression model are satisfied, under H_0 , the sampling distribution of t has a Student's t distribution with n-2 degrees of freedom

Hypothesis Tests Concerning β

Step 3: Find p-value. Compute sample statistic

$$t^* = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} / \sqrt{S_{xx}}}$$

– H_a: $\beta \neq \beta_0$ (two-sided test) $\quad \boxed{p\text{-value} = 2P(t > |t^*|)}$

– H_a: $\beta > \beta_0$ (one-sided test) $\quad \boxed{p\text{-value} = P(t > t^*)}$

– H_a: $\beta < \beta_0$ (one-sided test) $\quad \boxed{p\text{-value} = P(t < t^*)}$

P(t>|t*|), P(t>t*) and P(t<t*) can be found from the t table

Example: Age and Fatness

1. $\mathrm{H}_0 : \beta = 0, \quad \mathrm{H}_a : \beta \neq 0$

2. $$t^* = \frac{\hat{\beta} - 0}{\hat{\sigma} / \sqrt{S_{xx}}} = \frac{.55}{5.75 / \sqrt{2970}} = 5.21$$

   $\mathrm{df} = \mathrm{n} - 2 = 16$

3. $\text{p-value} < .005$

4. reject H_0

5. There is a significan t linear relationsh ip between age and fatness.

In Class Exercise

Writing a python code

-Make scattering plot of first and second columns
-Make a linear fit ($y= \alpha + \beta x$) to scattering plot and find α and β values and save it in *pdf* format.
-Make Regression Analysis by making ANOVA (An analysis of variance) table for first and second columns.