Linear Regression

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

Learning Objectives

After this lesson, you should be able to:

- Define simple linear regression
- Build a linear regression model using *statsmodels*
- Evaluate model fit using statistical analysis (t-tests, p-values, t-values, confidence intervals)

Here’s what’s happening today:

- Simple Linear Regression
  - Interpreting the regression’s coefficients
  - Are these coefficients significant?
  - Common Regression Assumptions
  - Model Fit and R^2
- The Normal Distribution
  - The 68 – 90 – 95 – 99.7 Rule
- Hypothesis Testing
  - Two-Tail Hypothesis Testing
  - t-values
  - p-values
  - Confidence Intervals
- The Student’s t-distribution

Simple Linear Regression

Simple Linear Regression

- The simple linear regression model captures a linear relationship between a single feature variable x and a response variable y

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

- y is the **response** variable (what we want to predict); also called *dependent* variable, *endogenous* variable, or *regressand*
- x is the **feature** variable (what we use to train the model); also called *explanatory* variable, *independent* variable, *exogenous* variable, or *regressor*
- β_0 and β_1 are the **regression's coefficients**; also called the *model's parameters*
  - β_0 is the line's intercept; β_1 is the line's slope
- ε is the **error** term; also called the residual

Simple Linear Regression (cont.)

- Given $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$, we can formulate the linear model as

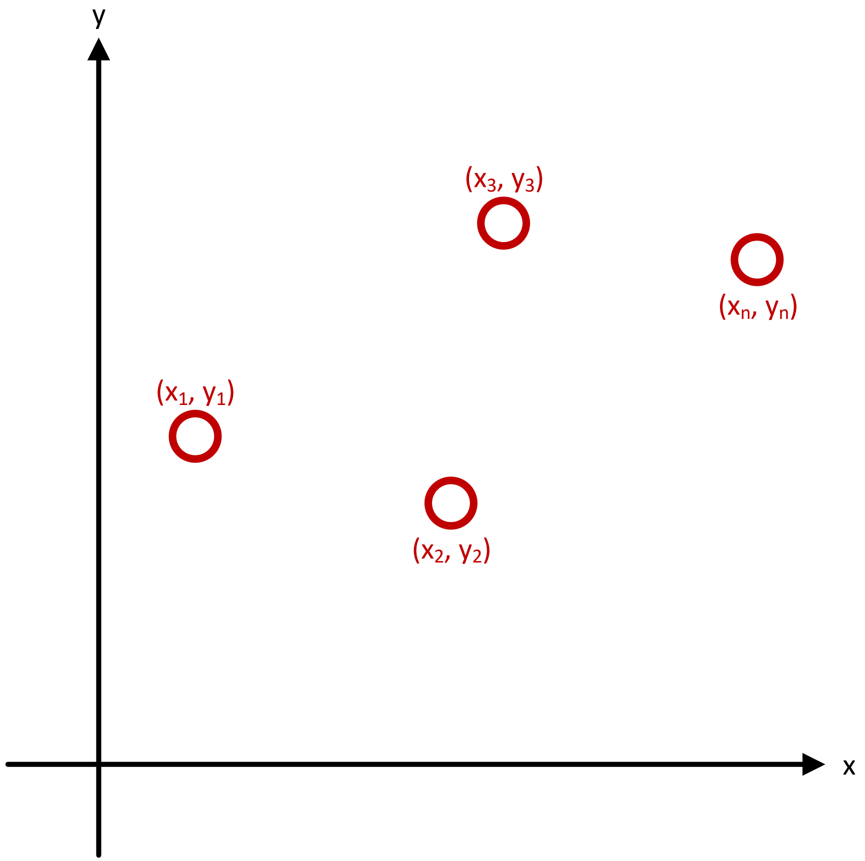
$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

- In words, this equation says that for each observation i, y_i can be explained by $\beta_0 + \beta_1 \cdot x_i$

- In our Python environment, x and y represent *pandas Series* and x_i and y_i their values at index i
- E.g. (SF housing dataset),
  - x is the property's size (`df.Size`)
  - y is the property's sale price (`df.SalePrice`)

Simple Linear Regression (cont.)

- ε_i is a "white noise" disturbance which **we do not observe**
  - ε_i models how the observations deviate from the exact slope-intercept relation
- **We do not observe** the constants β_0 or β_1 either, so we have to estimate them



Simple Linear Regression (cont.)

- Given estimates for the model coefficients $\hat{\beta}_0$ (β_0 hat) and $\hat{\beta}_1$ (β_1 hat), we predict y using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

- The hat symbol (^) denotes an estimated value

Simple Linear Regression (cont.)

- E.g. (SF housing dataset),

$$\widehat{SalePrice} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Size$$

How to interpret *statsmodels* report?

Dep. Variable:	SalePrice	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	297.4
Date:		**Prob (F-statistic):**	2.67e-58
Time:		**Log-Likelihood:**	-1687.9
No. Observations:	967	**AIC:**	3380.
Df Residuals:	965	**BIC:**	3390.
Df Model:	1		
Covariance Type:	nonrobust		

The model's fit

Is the model's fit significant?

The estimated coefficients
$\widehat{\beta}_0$ (the intercept) and
$\widehat{\beta}_1$ (the slope; "size")

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.1551	0.084	1.842	0.066	-0.010 0.320
size	0.749	0.043	17.246	0.000	0.664 0.835

Are these estimated significant?
(i.e., are they meaningful?; do they make sense?)

Omnibus:	1842.865	**Durbin-Watson:**	1.704
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3398350.943
Skew:	13.502	**Prob(JB):**	0.00
Kurtosis:	292.162	**Cond. No.**	4.40

Simple Linear Regression

Interpreting the regression's coefficients $\hat{\beta}$

Interpreting the regression's coefficients

$\hat{\beta}_0 = .155$	$\hat{\beta}_1 = .750$

- What's the unit of $\hat{\beta}_0$?
  - $[\hat{\beta}_0] = [\widehat{SalePrice}]$=\$M
- How to interpret $\hat{\beta}_0$?
  - $\hat{\beta}_0 = 0.155\ [\$M] = \$155k$
  - $\widehat{SalePrice}_{(Size=0)} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0$
  - The model predicts that a property of 0 sqft would cost \$155k

- What's the unit of $\hat{\beta}_0$?
  - $[\hat{\beta}_1] = \frac{[\widehat{SalePrice}]}{[Size]} = \frac{\$M}{1{,}000\ sqft}$
- How to interpret $\hat{\beta}_1$?
  - $\hat{\beta}_1 = \$.750/1{,}000\ sqft = \$750k/1{,}000\ sqft$
  - The model predicts that each additional 1,000 sqft costs buyers \$750k

Simple Linear Regression

Are the regression's coefficients $\hat{\beta}$ significant?

Are the regression's coefficients $\hat{\beta}$ significant?

The β coefficients follow a normal distribution:

$$\boldsymbol{\mu_{\beta} \sim N\left(\beta, \left(X^T X\right)^{-1} \sigma^2\right)}$$

(or)

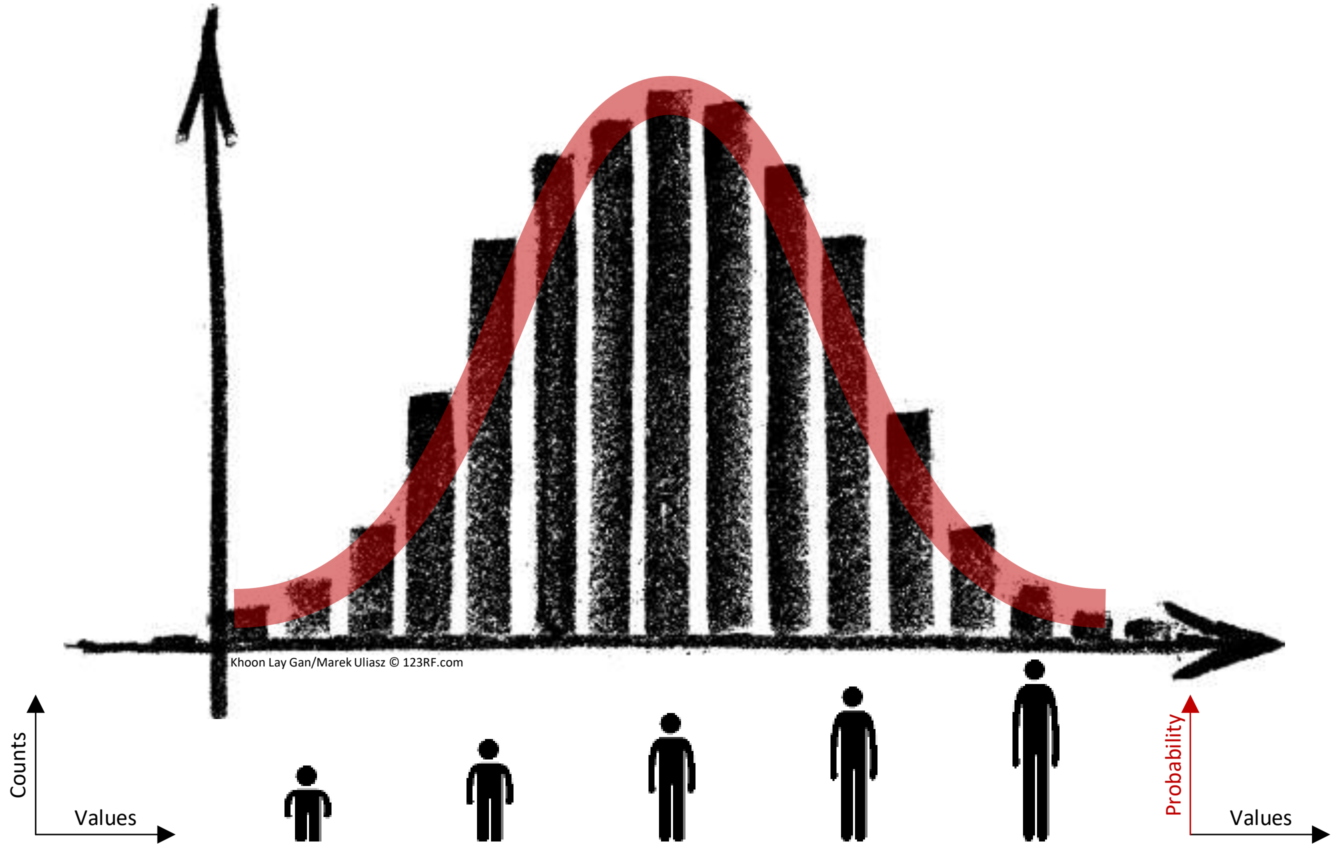
$$\boldsymbol{\mu_{\beta_j} \sim N\left(\beta_j, v_j \sigma^2\right)}$$

$$\left(X^T X\right)^{-1} = \begin{pmatrix} v_0 = v_{0,0} & \cdots & v_{0,j} & \cdots \\ \vdots & \ddots & & \\ v_{j,0} & & v_j = v_{j,j} & \\ \vdots & & & \ddots \end{pmatrix}$$

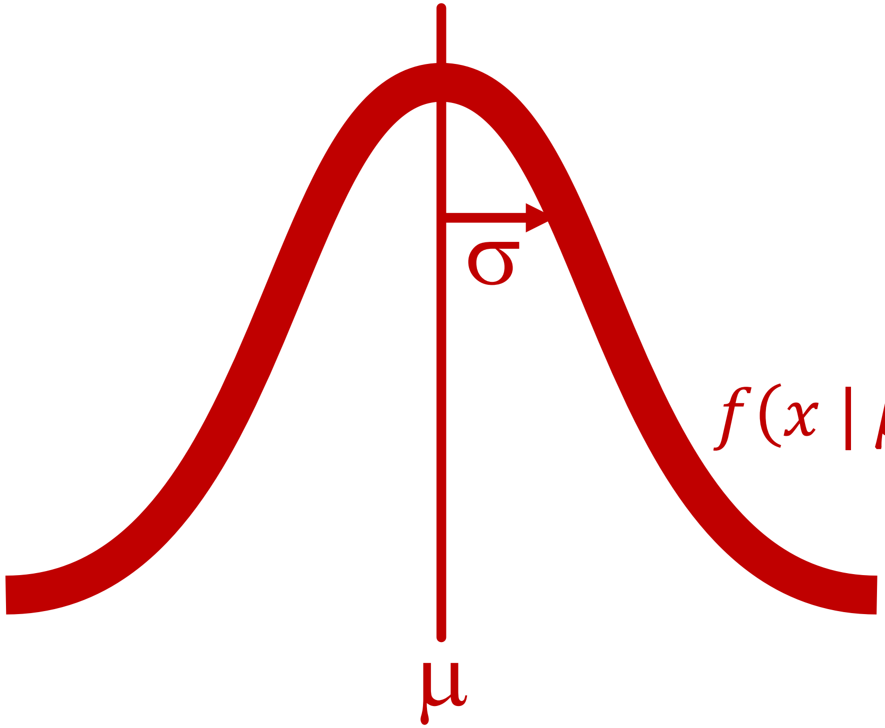
(v_j is the j^{th} diagonal element of $\left(X^T X\right)^{-1}$)

The Normal Distribution

People's height follows a bell shape distribution. (For men in the US, the average height is around 70 inches (5'10) with a standard deviation of 4 inches; few people are shorter than 67 inches; few are as tall as 73 inches)

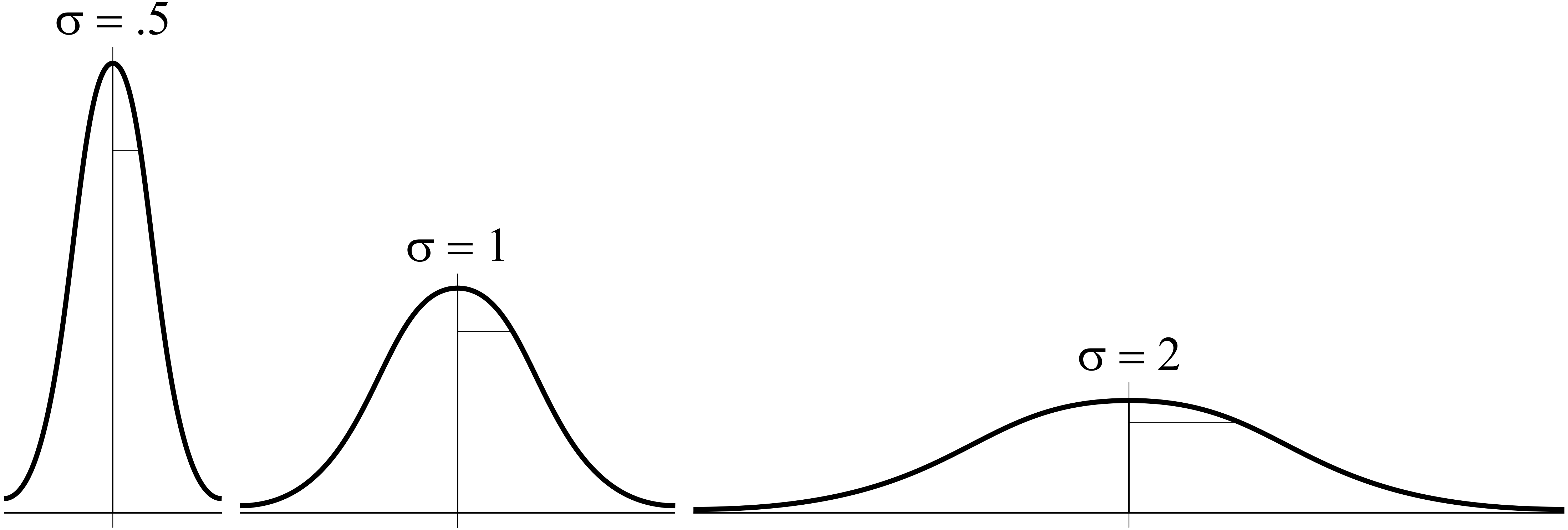


The Normal Distribution

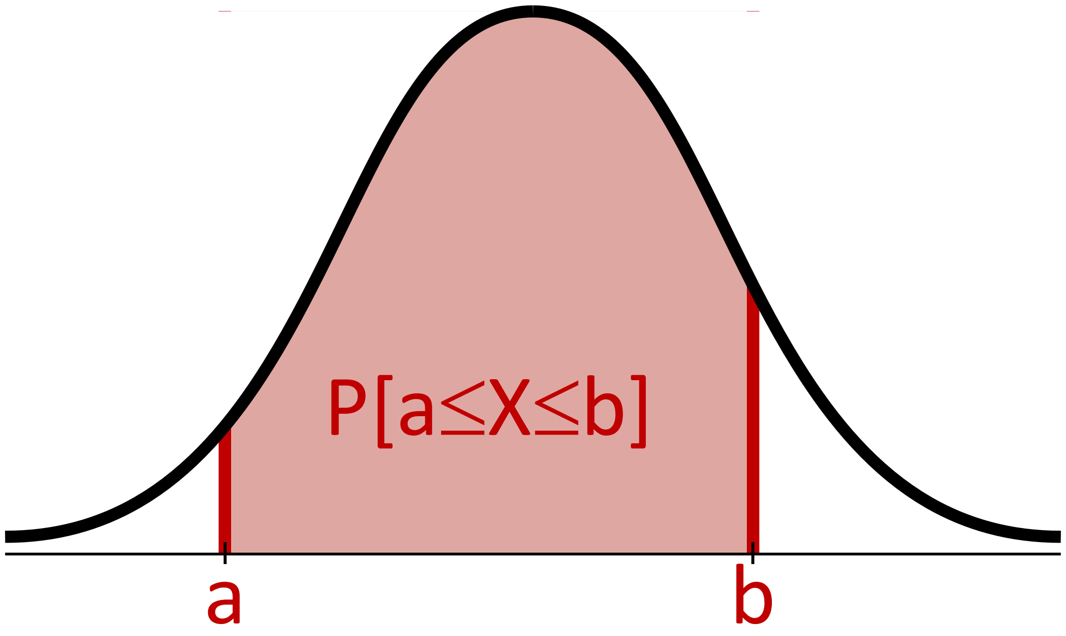


$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

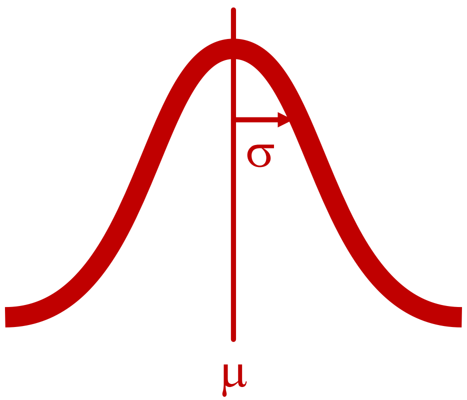
This bell-shaped curve is a probability density function (PDF): The area under the curve is always 1 (for any σ)

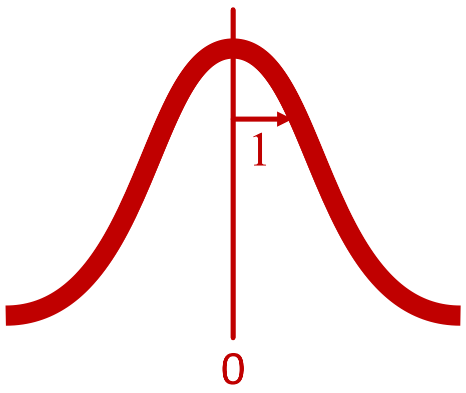


The area under the curve is called a cumulative distribution function (CDF)



The Standard Normal Distribution ($\mu = 0; \sigma = 1$)





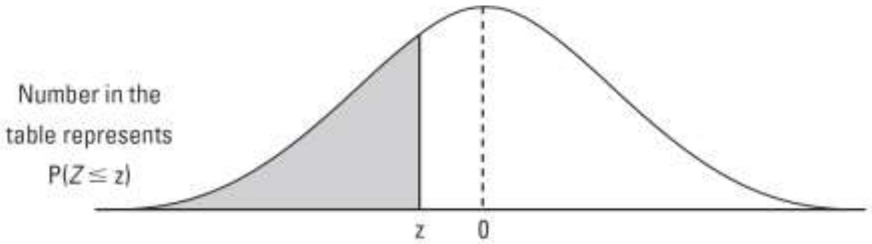
$$X \sim N(\mu, \sigma)$$

$$Z = \frac{X - \mu}{\sigma}$$

$$X = \mu + \sigma \cdot Z$$

$$Z \sim N(0,1)$$

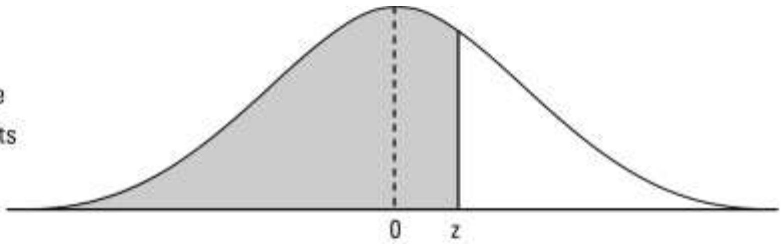
The Standard Normal Distribution Table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0003	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

The Standard Normal Distribution Table (cont.)

Number in the table represents $P(Z \leq z)$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

The 68 – 90 – 95 – 99.7 Rule

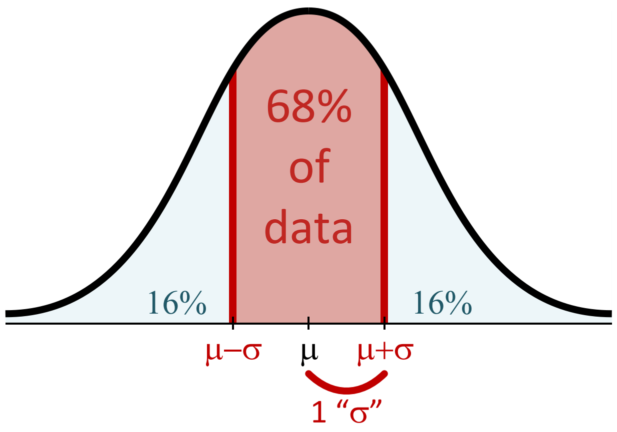
68%		
z	-1	1
$CDF(z)$	$.1587$	$.8413$
$P[-1 \leq Z \leq 1] = CDF(1) - CDF(-1)$ $= .8413 - .1587 = .6826 \cong 68\%$		

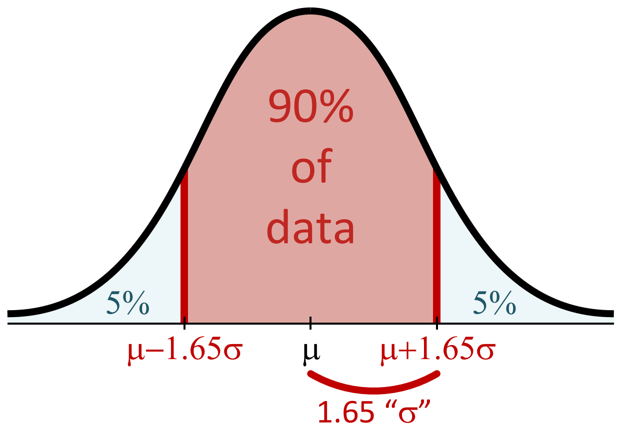
90%		
z	-1.65	1.65
$CDF(z)$	$.0495$	$.9505$
$P[-1.65 \leq Z \leq 1.65] = CDF(1.65) - CDF(-1.65)$ $= .9505 - .0495 = .9010 \cong 90\%$		

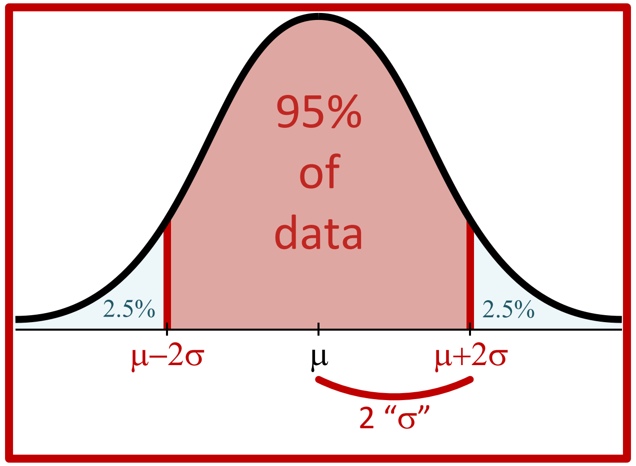
95%		
z	-2	2
$CDF(z)$	$.0228$	$.9772$
$P[-2 \leq Z \leq 2] = CDF(2) - CDF(-2)$ $= .9772 - .0228 = .9544 \cong 95\%$		

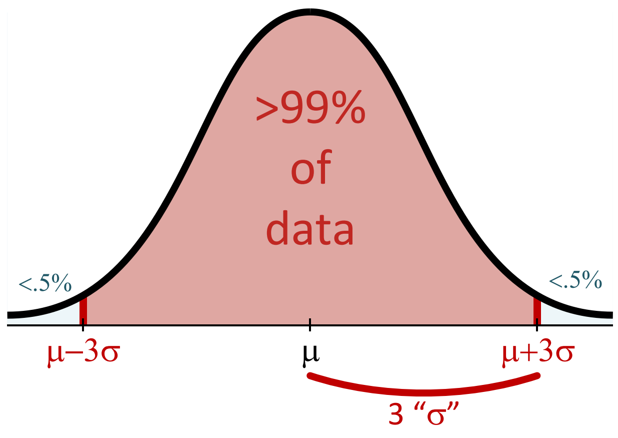
99.7%		
z	-3	3
$CDF(z)$	$.0013$	$.9987$
$P[-3 \leq Z \leq 3] = CDF(3) - CDF(-3)$ $= .9987 - .0013 = .9974 \cong 99.7\%$		

The 68 – 90 – 95 – 99.7 Rule (cont.)









Activity | The 68 – 90 – 95 – 99.7 Rule

EXERCISE

DIRECTIONS (10 minutes)

1. Adult women have an average height of 65 inches (5’5) and standard deviation of 3.5 inches. What are the lower and upper bounds for the middle 68%, 90%, 95%, and 99.7%?
2. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

Hypothesis Testing

Hypothesis Testing

- A hypothesis is an assumption about the a population parameter. E.g.,
  - μ_{β_0} = <a specific value, e.g. 0.155>
  - μ_{β_1} = <a specific value, e.g. 0.750>
- In both cases, we made a statement about a population parameter that may or may not be true
- The purpose of hypothesis testing is to make a statistical conclusion about **rejecting** or **failing to reject** such statement

Two-Tail Hypothesis Test

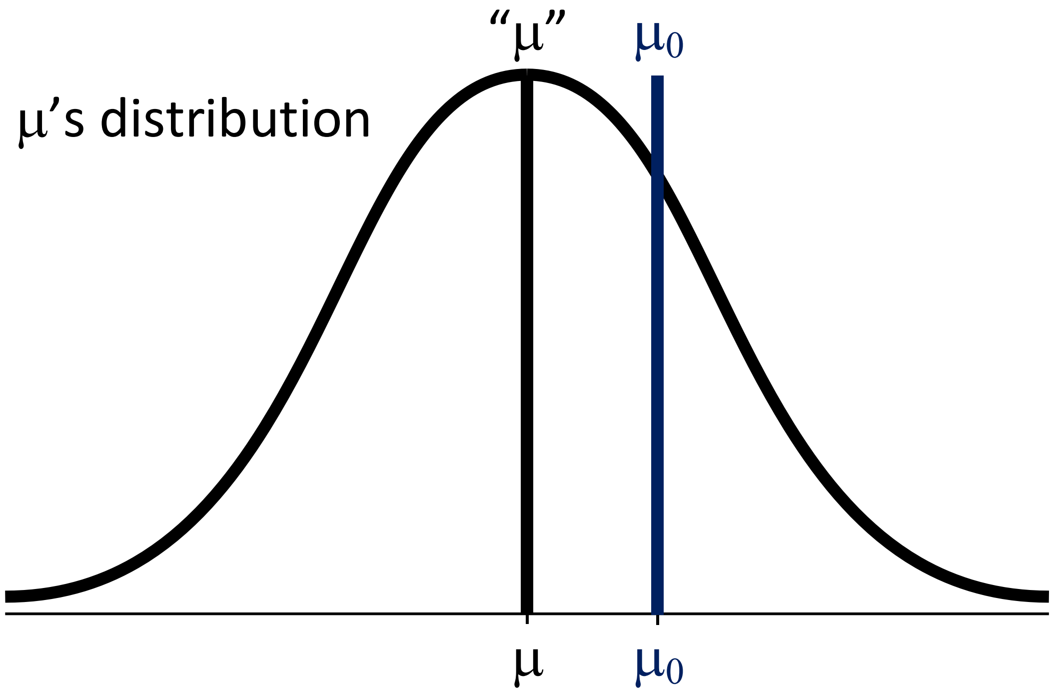
- The *null hypothesis* (H_0) represents the status quo; that the mean of the population is equal to a specific value:

$$H_0:\ \mu = \mu_0$$

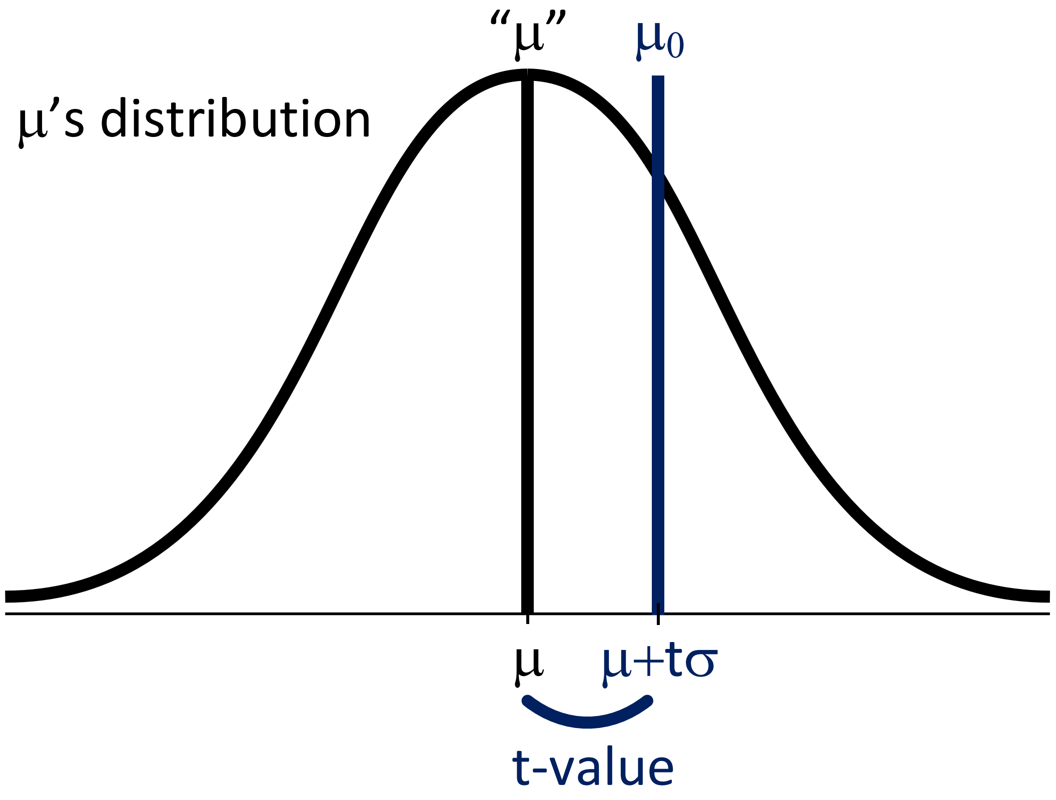
- The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis and holds true if the *null hypothesis* is found to be false:

$$H_a:\ \mu \neq \mu_0$$

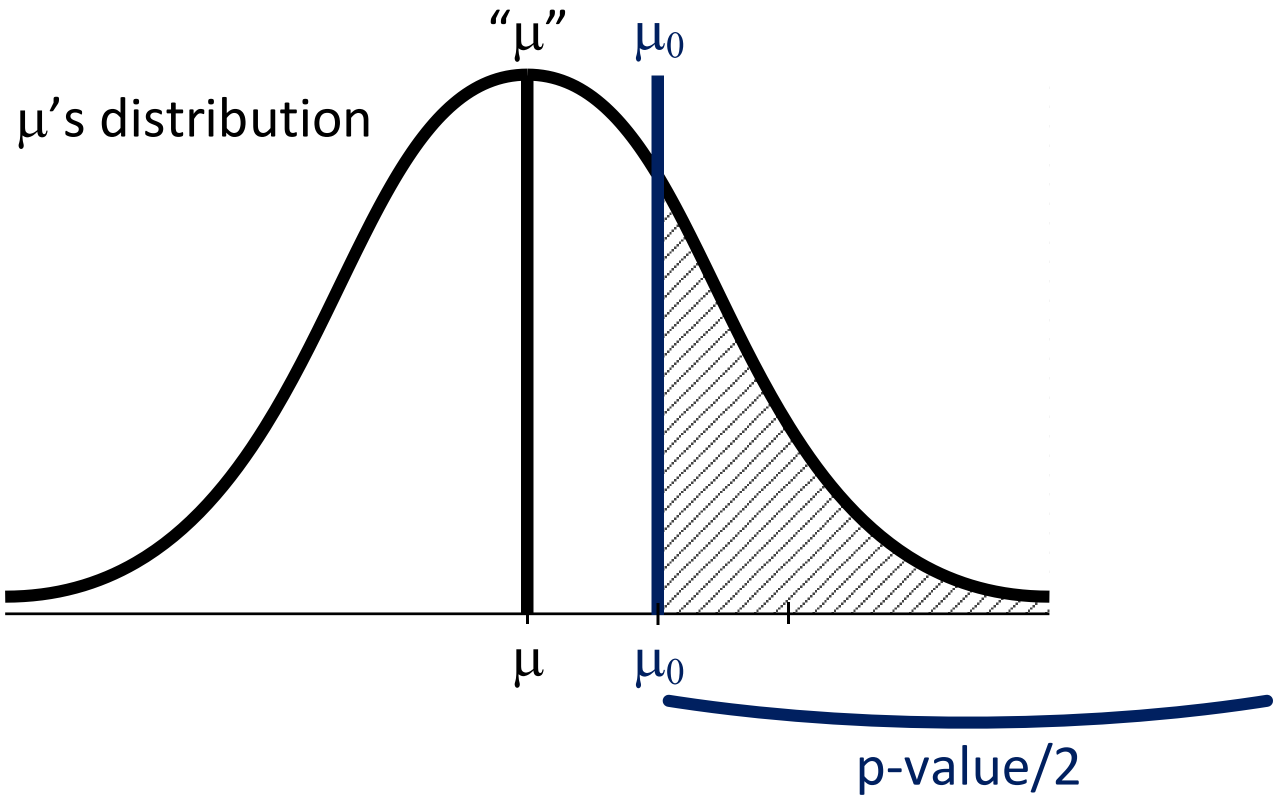
Two-Tail Hypothesis Test (cont.)



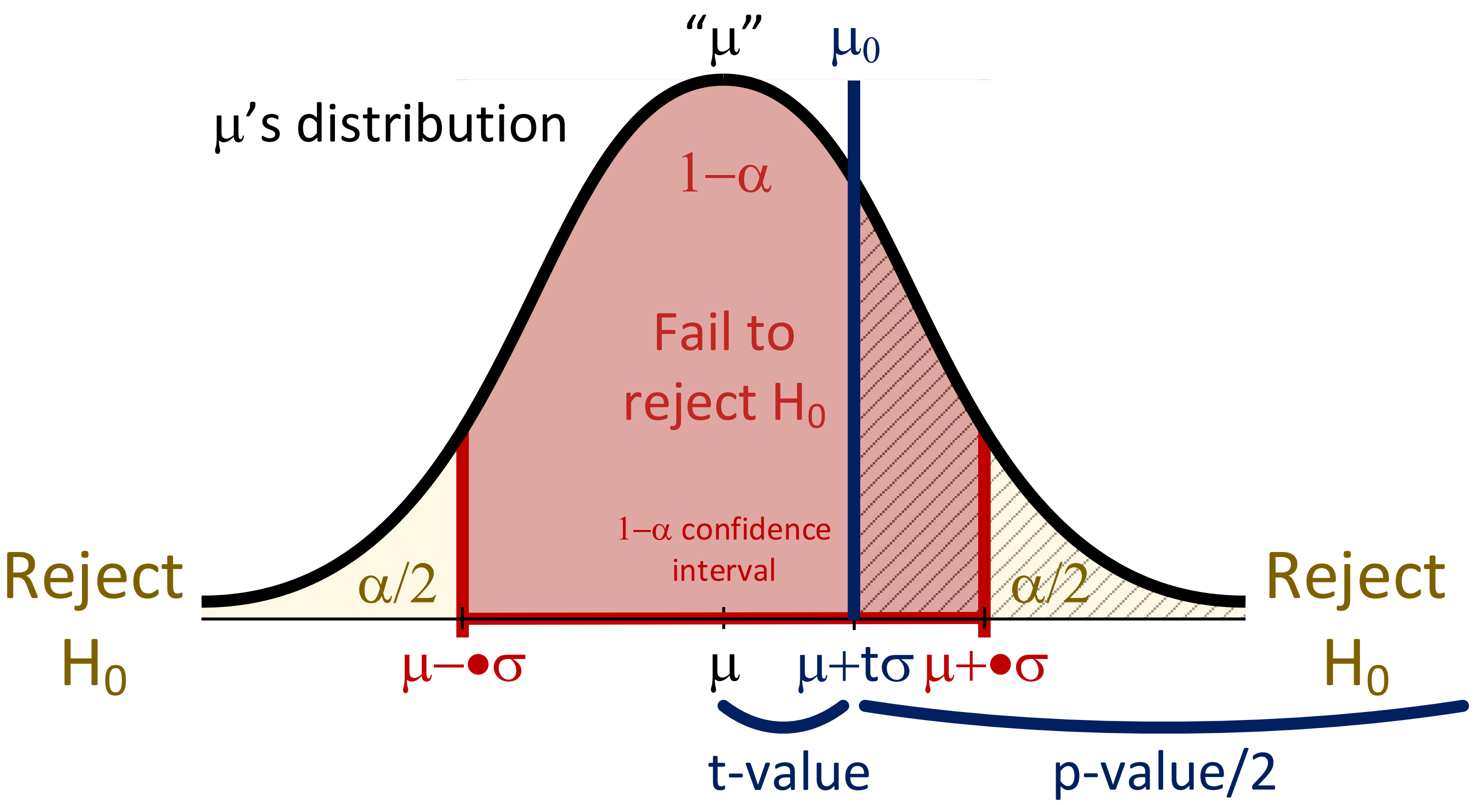
t-value measures the difference to μ_0 in σ. t-values of large magnitudes (either negative or positive) are less likely. The far left and right "tails" of the distribution curve represent instances of obtaining extreme values of t, far from μ



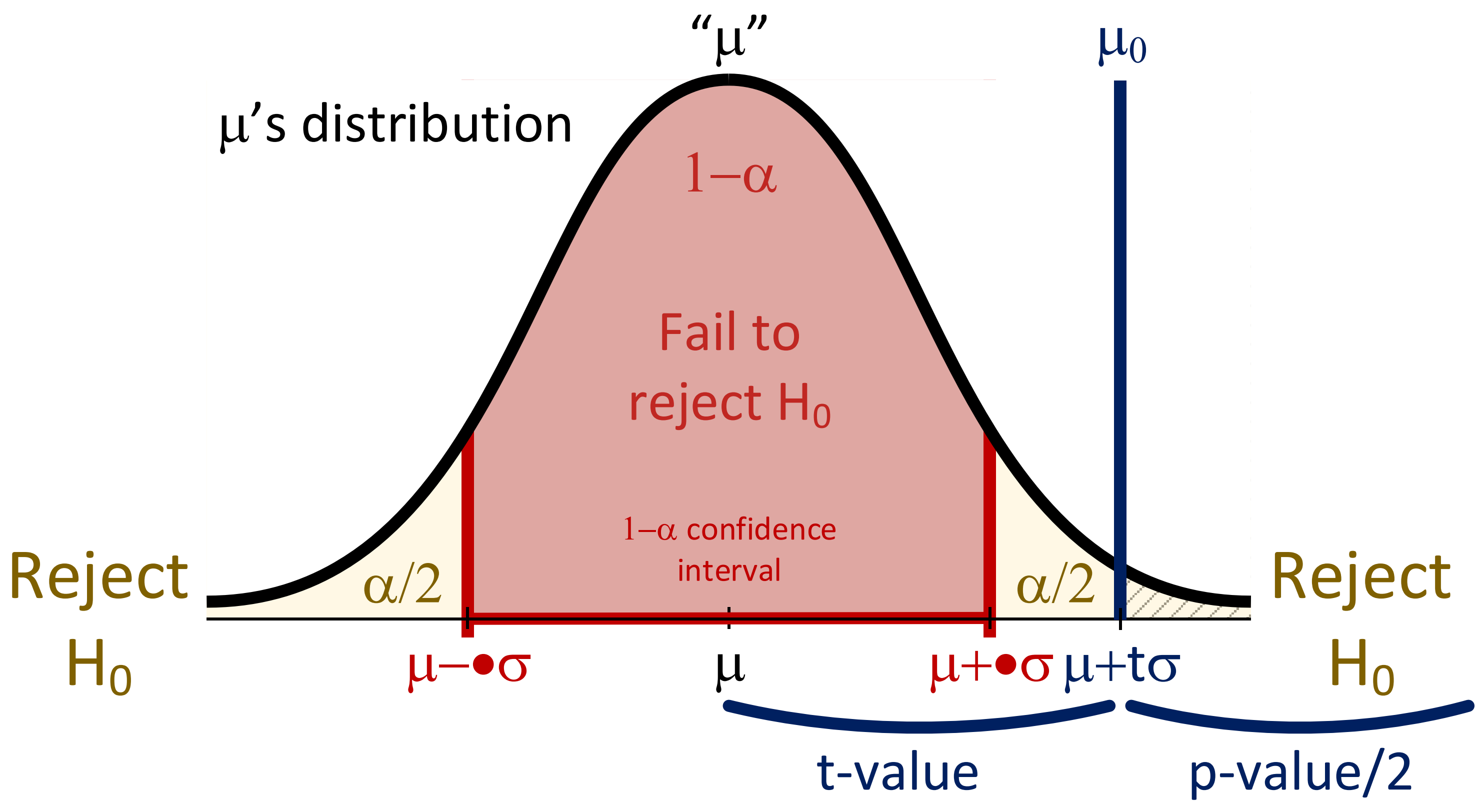
p-value determines the probability (assuming H_0 is true) of observing a more extreme test statistic in the direction of H_a than the one observed



Two-Tail Hypothesis Test *(simplified)* (cont.)



Two-Tail Hypothesis Test *(simplified)* (cont.)



Two-Tail Hypothesis Test (cont.)

|t-value|	p-value	$1 - \alpha$ Confidence Interval ($[\mu_0 -\cdot\, \sigma, \mu_0 +\cdot\, \sigma]$)	H_0 / H_a	Outcome
$< \cdot$	$> \alpha$	μ_0 is inside	Did not find evidence that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$ (assume)
$\geq \cdot$	$\leq \alpha$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$

Two-Tail Hypothesis Test ($\alpha = .05$) (cont.)

|t-value|	p-value	95% Confidence Interval ($[\mu_0 - 2\sigma, \mu_0 + 2\sigma]$)	H_0 / H_a	Outcome
$< "\sim 2"^{(*)}$ (*) (check t-table slide)	$> .05$	μ_0 is inside	Did not find evidence that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$ (assume)
$\geq "\sim 2"$	$\leq .05$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$

Simple Linear Regression

Are the regression's coefficients $\hat{\beta}$ significant? (cont.)

What β_1 would make our multiple linear regression model useless?

- (the simple linear regression model again, without intercept to keep things simple)

$$y = \beta_1 \cdot x + \varepsilon$$

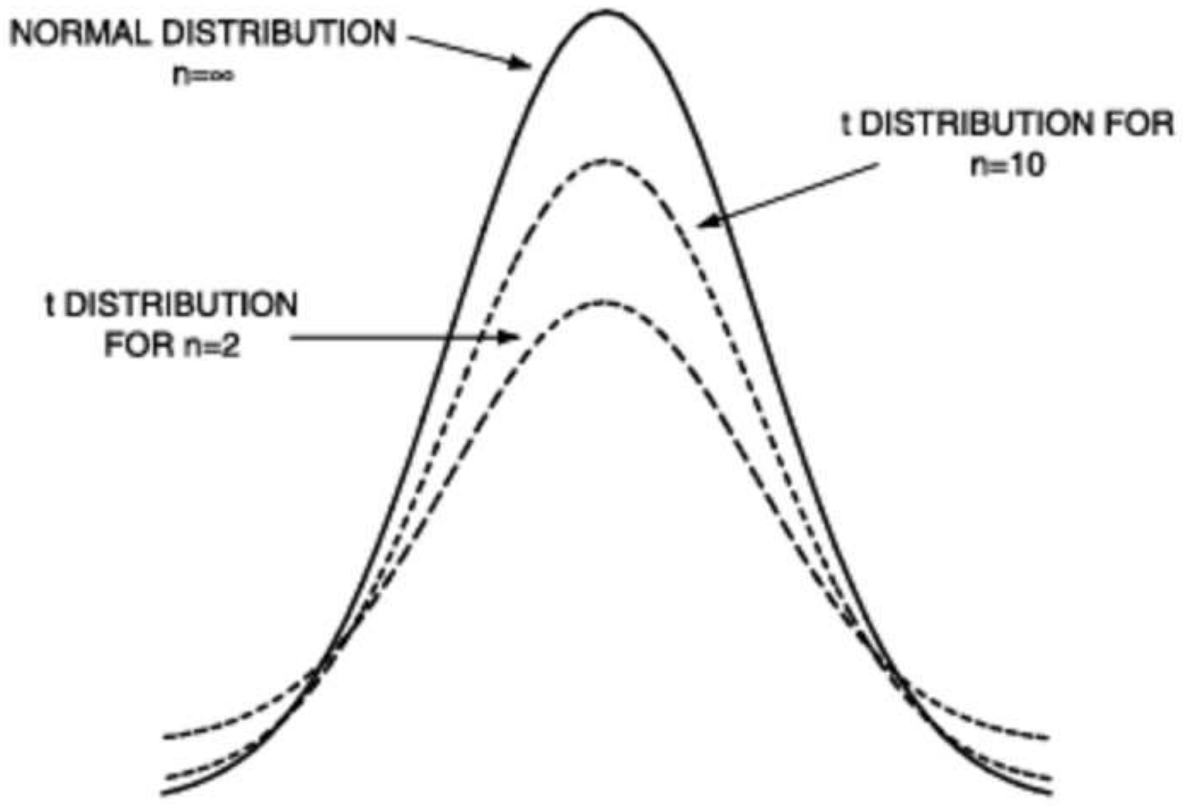
- Answer: If $\beta_1 = 0$, we don't have a linear model
  - ($y = 0$ isn't very exciting, is it?)

Details of *statsmodels*' coefficients table

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Feature variable j, e.g., "Intercept" or "Size"	$\widehat{\boldsymbol{\beta}}_{\mathbf{j}}$	$\sqrt{\mathbf{v_j}} \cdot \widehat{\boldsymbol{\sigma}}$	$z_j = \frac{\hat{\beta}_j}{\sqrt{v_j} \cdot \hat{\sigma}}$ (or) $\frac{coef}{std\ err}$	2 $\times$ area under the curve from the Student's t-distribution between $\|t\|$ and $+\infty$	$\hat{\beta}_j \pm z_{\alpha=.025} \cdot \hat{\sigma}$ (the value reported in the Student-t distribution table under the 5th column for $\alpha = .025$)

The Student’s t-distribution

FYI | We simplified things a bit… t-values refer to the Student's t-distribution, not the normal distribution; the reason behind this is that we substituted $\hat{\sigma}$ for σ (and $\hat{\beta}$ for β)



FYI | We simplified things a bit... t-values refer to the Student's t-distribution, not the normal distribution; the reason behind this is that we substituted $\hat{\sigma}$ for σ (and $\hat{\beta}$ for β) (cont.)

σ^2 is estimated by $\hat{\sigma}^2$

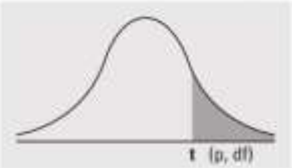
$$\hat{\sigma}^2 = \frac{1}{df} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

with

$$df = \underbrace{n}_{number\ of\ samples} - \underbrace{k}_{\substack{number\ of\ parameters \\ (intercept\ included)}}$$

Student's t-distribution table: as the sample size grows, the Student's t-distribution converges to a normal distribution

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	—	—	80%	90%	95%	98%	99%	99.9%

Simple Linear Regression

Are the regression's coefficients $\hat{\beta}$ significant? (cont.)

`SalePrice` as a function of `Size` (cont.)

Dep. Variable:	SalePrice	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	297.4
Date:		**Prob (F-statistic):**	2.67e-58
Time:		**Log-Likelihood:**	-1687.9
No. Observations:	967	**AIC:**	3380.
Df Residuals:	965	**BIC:**	3390.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.1551	0.084	1.842	0.066	-0.010 0.320
Size	0.7497	0.043	17.246	0.000	0.664 0.835

Omnibus:	1842.865	**Durbin-Watson:**	1.704
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3398350.943
Skew:	13.502	**Prob(JB):**	0.00
Kurtosis:	292.162	**Cond. No.**	4.40

$$SalePrice\ [\$M] = \underbrace{.155}_{\widehat{\beta}_0} + \underbrace{.750}_{\widehat{\beta}_1} \times Size\ [1{,}000\ sqft]$$

(the slope is significant but not the intercept)

`SalePrice ~ 0 + Size` (‘`0`’ meaning the intercept is forced to 0) (cont.)

Dep. Variable:	SalePrice	**R-squared:**	0.565
Model:	OLS	**Adj. R-squared:**	0.565
Method:	Least Squares	**F-statistic:**	1255.
Date:		**Prob (F-statistic):**	7.83e-177
Time:		**Log-Likelihood:**	-1689.6
No. Observations:	967	**AIC:**	3381.
Df Residuals:	966	**BIC:**	3386.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Size	0.8176	0.023	35.426	0.000	0.772 0.863

Omnibus:	1830.896	**Durbin-Watson:**	1.722
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3370566.094
Skew:	13.300	**Prob(JB):**	0.00
Kurtosis:	291.005	**Cond. No.**	1.00

$$SalePrice\ [\$M] = \underbrace{0.}_{\widehat{\beta}_0} + \underbrace{.810}_{\widehat{\beta}_1} \times Size\ [1{,}000\ sqft]$$

(the slope is significant)

`SalePrice ~ Size` (with outliers removed) (cont.)

Dep. Variable:	SalePrice	**R-squared:**	0.200
Model:	OLS	**Adj. R-squared:**	0.199
Method:	Least Squares	**F-statistic:**	225.0
Date:		**Prob (F-statistic):**	1.41e-45
Time:		**Log-Likelihood:**	-560.34
No. Observations:	903	**AIC:**	1125.
Df Residuals:	901	**BIC:**	1134.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.7082	0.032	22.152	0.000	0.645 0.771
Size	0.2784	0.019	15.002	0.000	0.242 0.315

Omnibus:	24.647	**Durbin-Watson:**	1.625
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	53.865
Skew:	0.054	**Prob(JB):**	2.01e-12
Kurtosis:	4.192	**Cond. No.**	4.70

$$SalePrice\ [\$M] =$$

$$\underbrace{.708}_{(was\ .155)} + \underbrace{.278}_{(was\ .750)} \times Size\ [1{,}000\ sqft]$$

(both intercept and slope are now significant)

DS

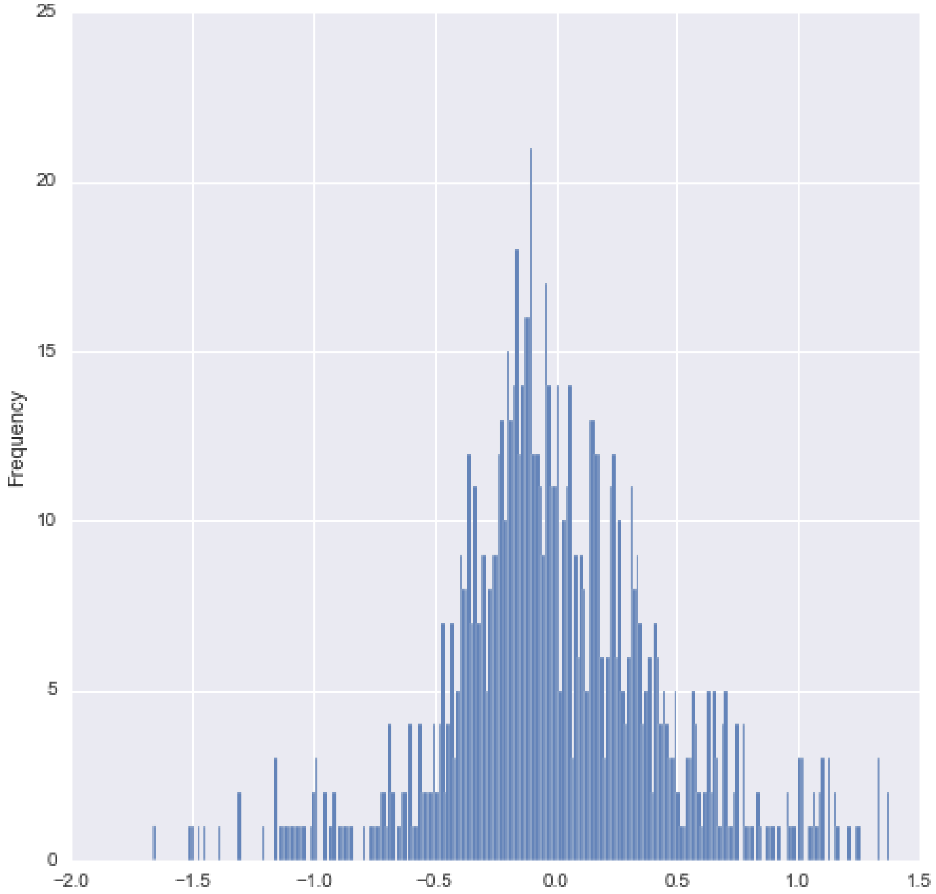
Simple Linear Regression

Common Regression Assumptions (part 1)

Common Regression Assumptions (part 1)

- The model is linear
  - x significantly explains y
- $\varepsilon \sim N(0,\ \cdot)$
  - Specifically, we expect ε to be 0 on average, i.e., $\mu_{\varepsilon} = 0$
- x and ε are independent
  - $\rho(x, \varepsilon) = 0$

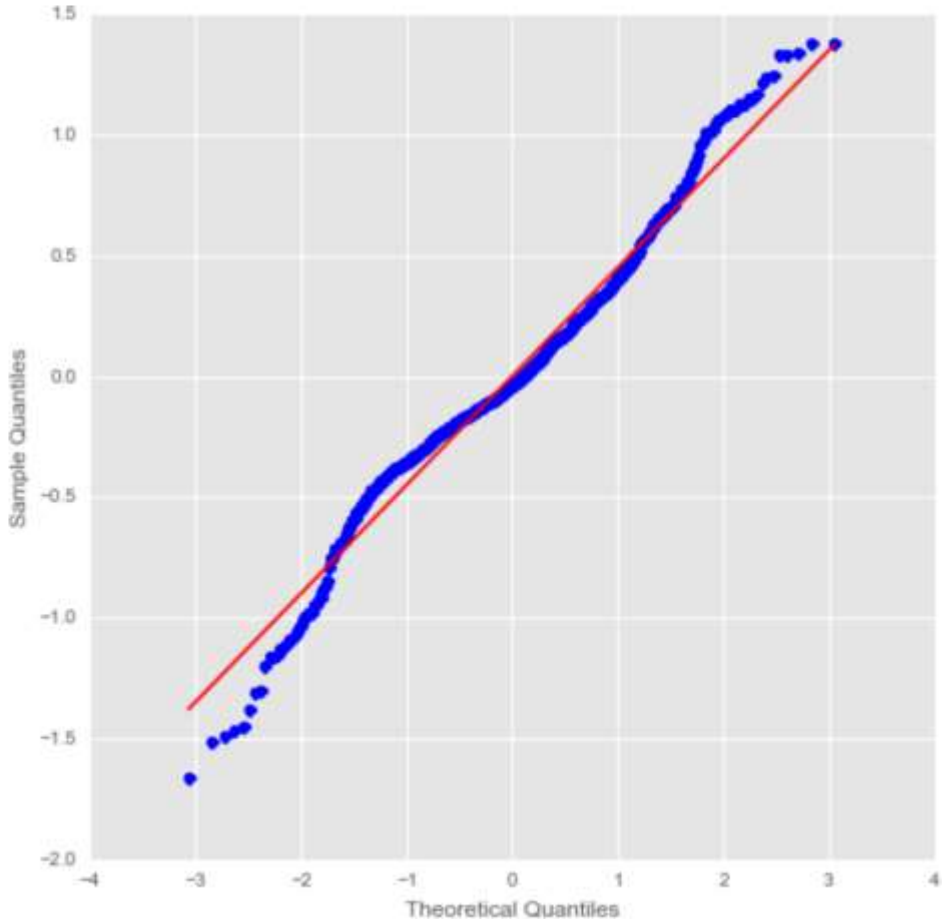
Is $\varepsilon \sim N(0, \ \cdot)$?



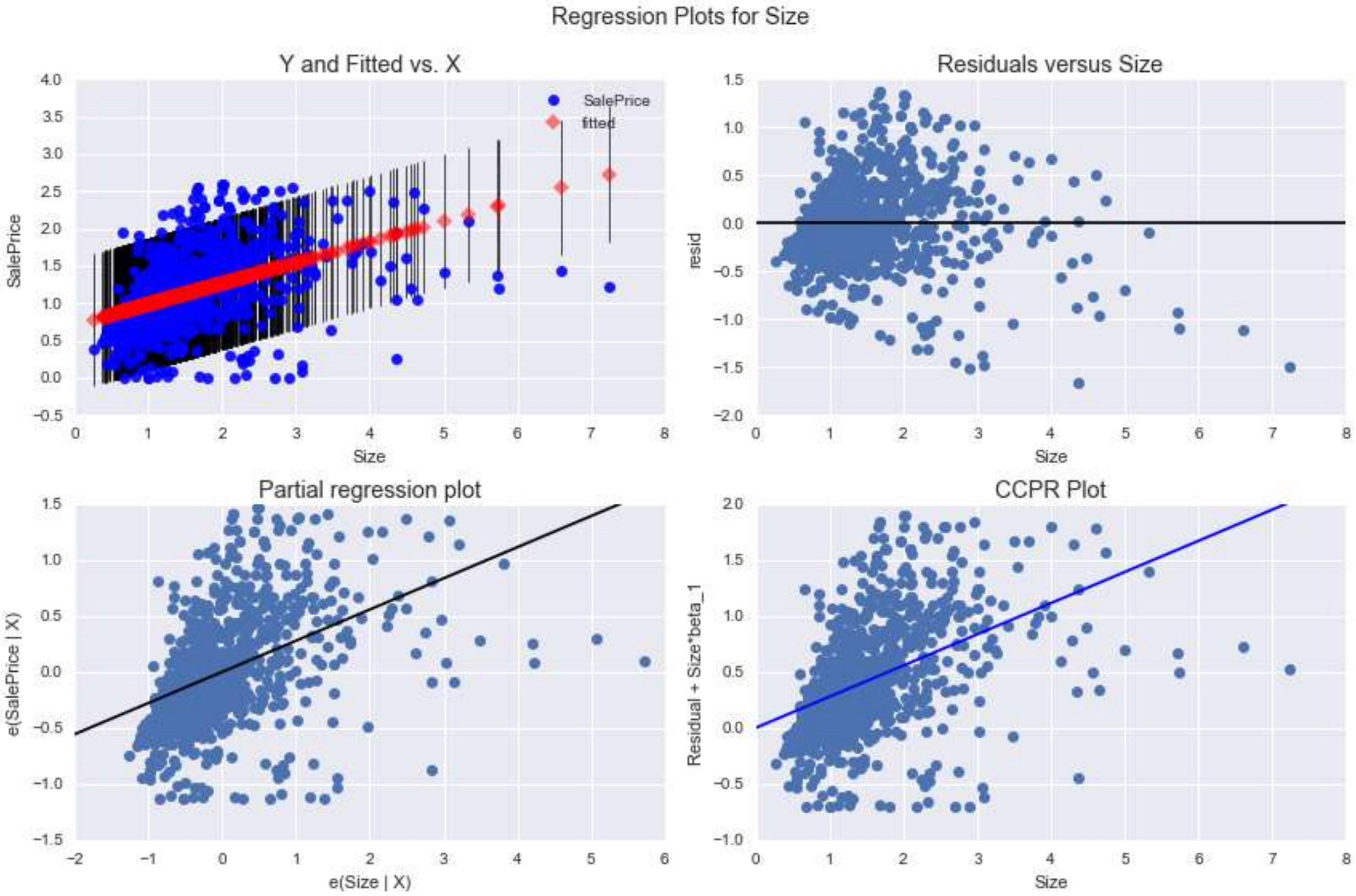
$\varepsilon \sim N(0, \cdot)$: `.qqplot()`

- "Quantile-Quantile (q-q) Plot"
- Graphical technique for determining if two datasets come from populations with a common distribution
- Plot of the quantiles of the first dataset (vertically) against the quantiles of the second's (horizontally)
- If unspecified, the second dataset will default to $N(0,1)$
- If the two datasets come from a population with the same distribution, the points should fall approximately along a 45-degree reference line
- The greater the departure from this reference line, the greater the evidence for the conclusion that the datasets have come from populations with different distributions

$\varepsilon \sim N(0,\ \cdot)$: `.qqplot()` (with `line = 's'`) (cont.)



x and ε are independent: `.plot_regress_exog()`



x and ε are independent: `.plot_regress_exog()` (cont.)

- Scatterplot of observed values (y) compared to fitted values ($\hat{y}$) with confidence intervals against the regressor (x)
  - `.plot_fit()`

- "Residual Plot"
  - Scatterplot of the model's residuals ($\hat{\varepsilon}$) against the regressor (x)

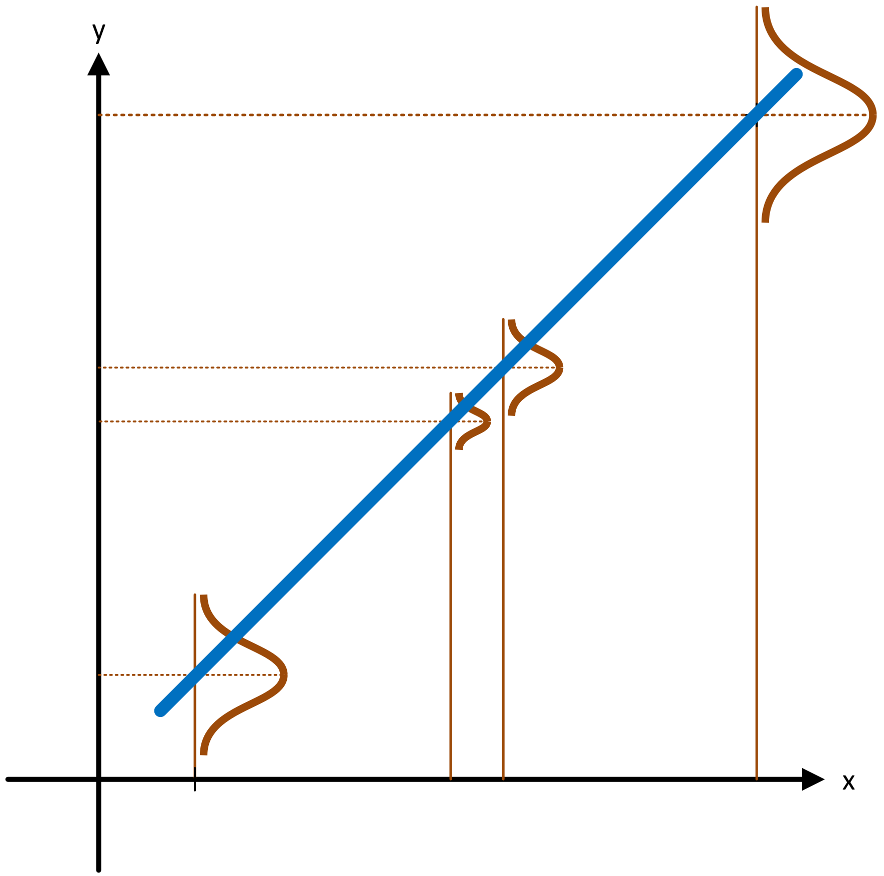
- "Partial Regression Plot" and "CCPR Plot (Component and Component-Plus-Residual)"
  - (useful for multiple regression)

Simple Linear Regression

Measuring the model's fit with R^2

Fit and Inference

- The deviations of the data from the best fitting line are normally distributed about the line. Since $\mu_\varepsilon = 0$, we "expect" that on average, the line will be correct
- How confident we are about how well the relationship holds depends on σ_ε^2



Measuring the model's fit with R^2

- When a measure of how much of the total variation in y, $\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_\varepsilon^2$, is explained by the portion associated with the explanatory variable x, $\sigma_{\hat{y}}^2 = \beta^2\sigma_x^2$; also called systematic variation (the variation explained by your model)

$$R^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} = \frac{\beta^2\sigma_x^2}{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}$$

- $0 \le R^2 \le 1$ (since $-1 \le \rho_{xy} \le 1$)

- $1 - R^2 = \frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2+\sigma_\varepsilon^2}$ is the idiosyncratic variation

  (the variation left unexplained by your model)

R^2: Goodness of Fit

When x significantly explains y	When x does not significantly explains y
❑ The fit is **better**	❑ The fit is **worse**
❑ The **explained** systematic variation dominates	❑ The **unexplained** idiosyncratic variation dominates
❑ σ_ε^2 is low (and/or $\beta^2\sigma_x^2$ is high)	❑ σ_ε^2 is high (and/or $\beta^2\sigma_x^2$ is low)
❑ $\boldsymbol{R^2} = \frac{1}{1+\underbrace{\frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2}}_{\cong 0}}$ **is closer to 1**	❑ $\boldsymbol{R^2} = \frac{1}{1+\underbrace{\frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2}}_{\gg 1}}$ **is closer to 0**