

MBC 638

LIVE SESSION WEEK 6

Agenda

Topic	Time	Sunday Section	Wednesday Section
Introduction	5 min	6:30-6:35	9:00-9:05
Quiz 2 Prep	40 min	6:35-7:15	9:05-9:45
Highlights from Week 6 Video	40 min	7:15-7:55	9:45-10:25
Review of Upcoming Assignments and Open Question	5 min	7:55-8:00	10:25-10:30

Quiz 2 Prep Question #3:Chi Square Test for Independence

Q3:Three work shifts producing the same product sorted the finished product into 4 categories based on its quality level and displayed the results in the following table. Determine whether there is dependence between the shift and the quality of the product? (Does product quality depend on the shift that produces it?) Assume alpha= 0.05

	1 st Shift	2 nd Shift	3 rd Shift	
Perfect product	185	175	170	<div>Ho: Quality and Shift are independent</div> <div>Ha: Quality and Shift are NOT independent</div>
Acceptable product	55	60	65	
Defective product	15	15	15	
Reworked product	10	15	15	

	Observed					Expected			
	1st shift	2nd shift	3rd shift	Total		1st shift	2nd shift	3rd shift	Total
Perfect product	185	175	170	530	Perfect product	176.67	176.667	176.67	530
Acceptable product	55	60	65	180	Acceptable product	60	60	60	180
Defective product	15	15	15	45	Defective product	15	15	15	45
Reworked product	10	15	15	40	Reworked product	13.333	13.3333	13.333	40
	265	265	265	795		265	265	265	795
	=CHISQ.TEST(H21:J22,N21:P22)								
	0.8403								

If P is low, Ho must go. P is .84 not lower than alpha =.05, so can NOT reject Ho – must accept Ho which says Quality and Shift ARE Independent.

At what alpha would you reject Ho.....anything larger than .84, so alpha = .85 you would reject Ho.

Quiz 2 Prep Question #4: Hypothesis Testing with Z statistics

4) A bullet manufacturer claims to have produced a projectile having a mean muzzle velocity of more than 3000 feet per second. From a random sample of 60 bullets he calculates a sample mean of 3012 feet per second and a sample standard deviation of 112 feet per second. Does the data from the sample support his claim?

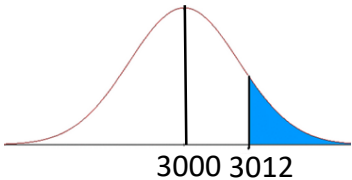
Solution:

Ho: $\mu \leq 3000$

n=60

Ha: $\mu > 3000$

1 sample test = purple chart



$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \qquad Z = \frac{3012 - 3000}{\frac{112}{\sqrt{60}}} = \frac{3012 - 3000}{\frac{112}{7.75}} = \frac{3012 - 3000}{14.45} = \frac{12}{14.45} = .830$$

Look up .830 in Z tables, =.7967 so 1-.7967=.2033

Or =NORMDIST(3012, 3000,14.459,TRUE) =.796712, 1-.796712=.203288

If p is low then how must go, but in this case alpha is not provided...however p is not lower than the typical alpha of .05, if this is what you assumed then You can't reject Ho which means that the sample does NOT support his claim.

One-tail test	
Left-tail	Upper/right-tail
$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

Sample size	
Large	μ_0
$n \geq 30$	
(or σ known)	

Test statistic	
$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	μ_0
Can replace s with σ if known	

Upper/right-tail
$p = \text{area right of } Z \text{ or } t$

Table C Standard normal distribution (continued)[illegible]

Highlights: Video Segment 3.7: Hypothesis Testing

Use sample data from a population to confirm or reject a statement that we make about the population.

Hypothesis Statements can tell us if two sets of data are really different from each other, or from a standard value.

Also provides the probability of being right or wrong and the risk of making the wrong decision.

Only pertains to population parameters – mean and standard deviation.

Uses of Hypothesis Testing

- Hypothesis testing can tell us:
 - If two sets of data are really different
 - If population parameter varies from a standard
 - Probability of being right or wrong
 - Risk of making an incorrect decision

Null hypothesis	Alternative hypothesis
$H_0: \mu \text{ or } \sigma = (\text{or } \leq, \text{ or } \geq) \text{ a number}$	$H_a: \mu \text{ or } \sigma \neq (\text{or } <, \text{ or } >) \text{ a number}$
Captures "other" results	Captures results of interest
Locus of equality condition	$H_a: \mu \neq 10$
$H_0: \mu = 10$	$H_a: \mu > 10$
There is <i>no</i> difference!	There <i>is</i> a difference!

The alternative:
Is what you want the result to be....we want to have improved our process, reduced the cycle time, made it lower.

Quiz 2 Prep Additional Question: Hypothesis Testing with T statistics

The gas industry has issued data that the standard cost of gas is \$2.14 per gallon at stand alone gas stations. You believe the cost of gas at supermarket gas stations is greater than the cost of gas at stand alone gas stations.

You collect data from 10 signs for super market gas stations. Assume an alpha of .05.

The gas prices at the super market gas stations are (\$/gallon): 2.04, 2.05, 2.05, 2.09, 2.09, 2.07, 2.10, 2.12, 2.05, 2.06

Is the cost of gas at supermarket gas stations greater than the gas at stand alone gas stations?

Ho: super market gas prices <= standard gas price at stand alone stations
Ha: super market gas prices > standard gas price at stand alone stations

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$
$$df = n - 1$$

$$\frac{2.072 - 2.14}{\frac{.027}{\sqrt{10}}} = \frac{2.072 - 2.14}{\frac{.027}{\sqrt{10}}} = \frac{-.068}{\frac{.027}{3.16}} = \frac{-.068}{.0085} = \frac{-.068}{.0085} = -8$$

$$df = n - 1 = 10 - 1 = 9$$

Upper/right-tail

$H_o: \mu \leq \mu_o$

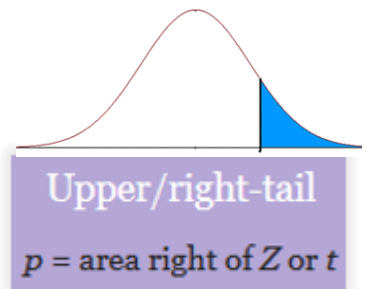
$H_a: \mu > \mu_o$

Small

$n < 30$

(or σ unknown)

Lookup in the t table, -8 at df=9, note there is no negative in the table the probabilities are symmetrical, so < than .005
Or = T.DIST(-8, 9, TRUE) in excel = 1.10674E-05, so almost 0
We want the area to the right, so 1- .00001 = almost 1



If p is low, Ho must go, P of almost 1 is NOT lower than alpha = .05, so can NOT reject
Ho, you must accept that super market gas prices are lower than the stand alone gas station prices.

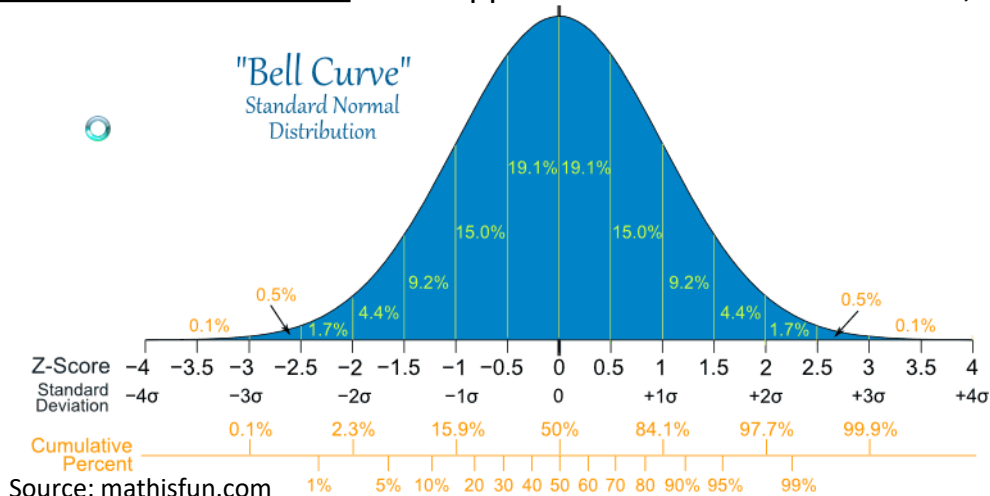
You can not conclude that super markets have higher prices than stand alone gas stations.

Table D *t*-Distribution

		Confidence level				
		80%	90%	95%	98%	99%
		Area in one tail				
		0.10	0.05	0.025	0.01	0.005
		Area in two tails				
		0.20	0.10	0.05	0.02	0.01
df	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	31	1.309	1.696	2.040	2.453	2.744
	32	1.309	1.694	2.037	2.449	2.738
	33	1.308	1.692	2.035	2.445	2.733
	34	1.307	1.691	2.032	2.441	2.728
	35	1.306	1.690	2.030	2.438	2.724
	36	1.306	1.688	2.028	2.435	2.719
	37	1.305	1.687	2.026	2.431	2.715
	38	1.304	1.686	2.024	2.429	2.712
	39	1.304	1.685	2.023	2.426	2.708
	40	1.303	1.684	2.021	2.423	2.704
	50	1.299	1.676	2.009	2.403	2.678
	60	1.296	1.671	2.000	2.390	2.660
	70	1.294	1.667	1.994	2.381	2.648
	80	1.292	1.664	1.990	2.374	2.639
	90	1.291	1.662	1.987	2.368	2.632
	100	1.290	1.660	1.984	2.364	2.626
	1000	1.282	1.646	1.962	2.330	2.581
	z	1.282	1.645	1.960	2.326	2.576

Highlights: Video Segment 3.4: Normal(Continuous Data)

Total Area under the curve = 1 or 100% of the opportunities are under the curve, since they are probabilities



For Normal, In order to solve for the probability for continuous data....

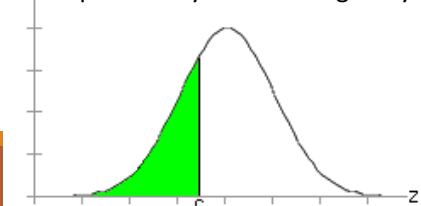
1. Calculate the standard value for Z in order to look up the probability

$z = \frac{x - \mu}{\sigma}$, μ = average of the sample, σ = standard deviation, x = the point you are interested in finding the probability for

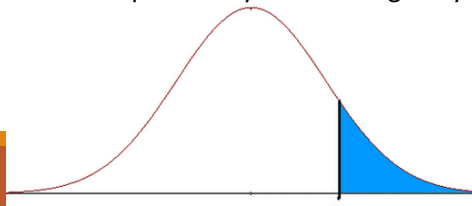
2. Solve for Z, then look it up in the table to convert to a probability.

Remember: The Z value is always for everything to left of the point you are looking for the probability of....

Use the probability the Z table gives you



Subtract the probability the Z table gives you from 1



Find the probability for each point and subtract them



Highlights: Video Segment 3.4: Normal(Continuous Data)

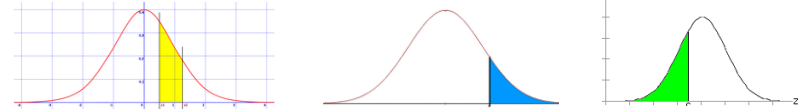
Using Excel to solve for probabilities and Z values using the normal distribution

1 – If you didn't calculate a Z value first

`=norm.dist(60,50,5,True)` = .97725 or 97.725%

Means give me the probability if I want to know what probability there is a child measures shorter than 60 inches, if the average is 50, and the standard deviation is 5, true means I want everything to the left of the curve.

DRAW the PICTURE – so you know what probability you are looking for.



2 – If you already have or calculated the Z value

`=norm.s.dist(2,true)` = .97725 or 97.725%

Means give me the probability for a Z test statistic of 2, and true means that I want everything to the left of the curve.

3 – If you have a probability and want a z value

`=norm.s.inv(.97725)`

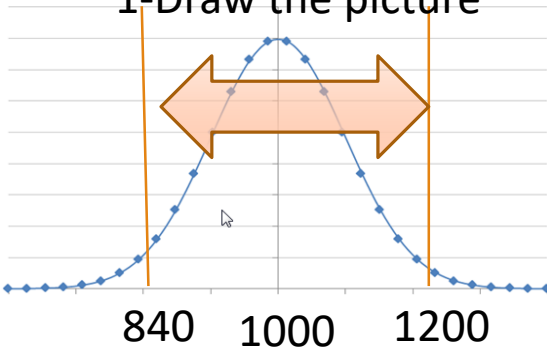
Means give me the Z value associated with the probability of .97725.

Quiz 2 Prep Question 1: Practice with Z calculations

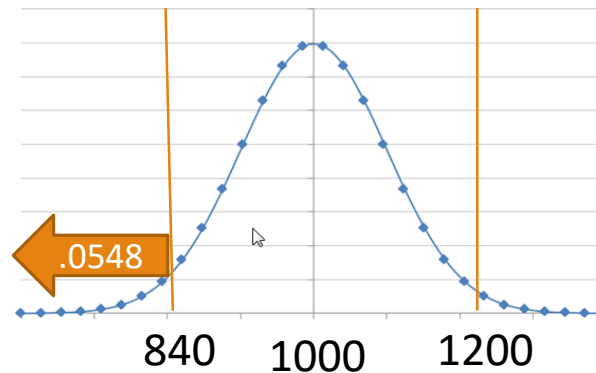
The distribution of weekly incomes of supervisors at the ABC Company follows the normal distribution, with a mean of \$1000 and a standard deviation of \$100.

What percent of the supervisors have a weekly income between \$840 and \$1200?

1-Draw the picture



2-Think about what you are calculating related to the picture



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

$$Z = \frac{840 - 1000}{100} = -1.6$$

Look up in tables, $p = .0548$

Or in Excel

`=NORM.DIST(840,1000,100,TRUE)`

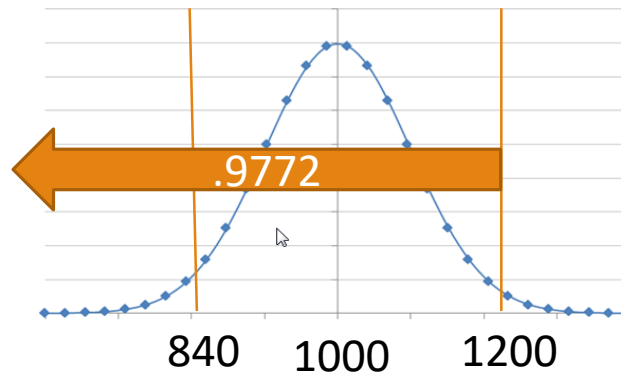
$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{1200 - 1000}{100} = 2$$

Look up in tables, $p = .9772$

Or in Excel

`=NORM.DIST(1200,1000,100,TRUE)`



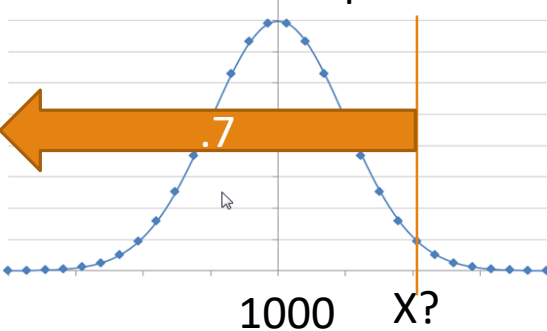
$.9772 - .0548 = .9224$, so 92.24% have a weekly income between \$840 and \$1200

Table C Standard normal distribution (continued)

Quiz 2 Prep Question 1 Extended: Practice with Z calculations

The distribution of weekly incomes of supervisors at the ABC Company follows the normal distribution, with a mean of \$1000 and a standard deviation of \$100. Management wants to give bonuses to those supervisors within the top 30% of weekly incomes. What is the weekly income cut off, the lowest weekly income a supervisor can have and still receive the bonus?

1-Draw the picture



2-Think about what you are calculating related to the picture

$$Z = \frac{X - \mu}{\sigma} \quad Z = \frac{x - 1000}{100} = ?$$

We know $p = .7$, what Z goes with that p value...look it up from the table = .52 or in Excel=NORM.S.INV(0.7)=.52

$$.52 = \frac{x - 1000}{100}$$

$$.52 * 100 = X - 1000$$

$$52 = X - 1000$$

$$52 + 1000 = X$$

$$1052 = X$$

\$1052 is the lowest weekly income a supervisor can have and still receive a bonus

Quiz 2 Review

Highlights: Video Segment 3.5: Binomial(Discrete Data)

No set shape
Mean: $\mu=np$, n = number of trials, p =probability of success
Variance: $\sigma^2= n \times p(1-p)$



A binomial experiment is an experiment which satisfies these four conditions:
A fixed number of trials
Each trial is independent of the others
There are only two outcomes
The probability of each outcome remains constant from trial to trial.



You're taking a quiz with five true/false questions. You didn't study and plan to guess.
What's the probability you get three questions correct?

Find $P(X = 3)$, the probability that the number of successes is equal to three.

- $n = 5$
- $p = 0.5$

Example: Binomial Table

		p (probability of a success)						
n	X	0.10	0.15	0.20	...	0.40	0.45	0.50
...			
4	0	0.6561	0.5220	0.4096		0.1296	0.0915	0.0625
	1	0.2916	0.3685	0.4096		0.3456	0.2995	0.2500
	2	0.0486	0.0975	0.1536		0.3456	0.3675	0.3750
	3	0.0036	0.0115	0.0256		0.1536	0.2005	0.2500
	4	0.0001	0.0005	0.0016		0.0256	0.0410	0.0625
5	0	0.5905	0.4437	0.3277	...	0.0778	0.0503	0.0312
	1	0.3280	0.3915	0.4096		0.2592	0.2059	0.1562
	2	0.0729	0.1382	0.2048		0.3456	0.3369	0.3125
	3	0.0081	0.0244	0.0512		0.2304	0.2757	0.3125

Use the tables in the back of the book or Excel to calculate probabilities for a binomial distribution for discrete data or use Excel:
`=binom.dist(3,5,.5,False)` = .3125

Means give me the probability that I get 3 successes, out of 5 trials, when each has a probability of success of .5, and False means exactly 3 successes(True would mean all those probabilities up to and including 3 successes:1,2,3 successes all added together)

Quiz 2 Review

Highlights: Video Segment 3.5: Binomial(Discrete Data)

Another Example:

For a multiple choice test that you are guessing on, you want to know the probability you get at least 3 correct, on a test that has 5 multiple choice questions, and each has 4 choices.

$n=5$ test questions, $p=.25$ (chance of answering correctly on each problem($1/4$)), $x \geq 3$

`=BINOM.DIST(3,5,0.25,FALSE)`

This formula means, the probability that I get 3 questions correct, with 5 questions on the test, and 4 answers for each question so a $1/4=.25$ chance of getting each correct, and false means I don't want the cumulative percent because I won't pass the test if I get 1 or 2 correct. This will give me probability of getting exactly 3 correct, then I would do the same with the probability at 4 and 5 correct and add the three probabilities together – because I want to know the probably of getting at least 3 correct, which means 3 or more correct.

`=BINOM.DIST(3,5,0.25,FALSE)= .088`

`=BINOM.DIST(4,5,0.25,FALSE)= .015`

`=BINOM.DIST(5,5,0.25,FALSE)= .0010`

.104 or 10.4% probability that you get at least 3 correct

Or `=BINOM.DIST(2,5,0.25,TRUE)` means probability I get 1 or 2 correct, then subtract from 1 to get probability of 3,4,5 correct

`=.896`, so $1-.896=.1035$ or .104, so 10.4% probability that you get at least 3 correct

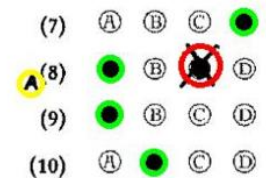


Table B Binomial distribution

<i>n</i>	<i>X</i>	<i>p</i>								
		0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
	2	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
	2	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
	3	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
	2	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
	3	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
	4	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0312
	1	0.3280	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1562
	2	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
	3	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
	4	0.0004	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1562
	5		0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0312
6	0	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
	2	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
	3	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
	4	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
	5	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
	6			0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
	2	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
	3	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
	4	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
	5	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
	6		0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
	7				0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0312
	2	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
	3	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
	4	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
	5	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
	6		0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
	7			0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
	8					0.0001	0.0002	0.0007	0.0017	0.0039

Note: Blank entries indicate a binomial probability of less than 0.00005.

Quiz 2 Prep Question

Twenty percent of the employees of ABC Company use direct deposit and have their wages sent directly to the bank. Assume we random sample five employees. What is the probability that all five employees use direct deposit?

Solution:

$n=5$, trials is 5

$p=.2$, 20% chance they use direct deposit(yes or no)

$X=5$, we want to know the probability 5 employees used direct deposit

=Table B in your book, $n=5, x=5, .2$ column = **.0003 is the probability all 5 employees use direct deposit**

Or in Excel =BINOM.DIST(5,5,0.2,FALSE)

BINOM.DIST(successes, trials, probability, cumulative), false because we want exactly 5 successes not cumulative)

Highlights: Video Segment 5.4 Sample Size for Continuous

Sample Size Formula for Continuous Data

$$n = \left(\frac{z^* \hat{\sigma}}{E} \right)^2$$

Look at sample size formula:
higher error = smaller sample size
Higher z value (higher confidence level) = larger sample size
Higher sigma (std deviation) = larger sample size

Example: Time to Complete Job

- Suppose you have collected a simple random sample of data and found the standard deviation to be three minutes.
- How many samples are needed to detect a change in job completion time after a process improvement project is implemented?
 - You are okay with a margin of error of two minutes.
 - Assume you want 95% confidence.

Example: Time to Complete Job (cont.)

$$n = \left(\frac{1.96(3)}{2} \right)^2$$
$$= 8.6 \approx 9$$

- z^* at 95% confidence = 1.96
- $\hat{\sigma} = 3$
 - Estimated population standard deviation
 - Equivalent to sample standard deviation, s
- $E = 2$

Highlights: Video Segment 1.4 Types of Data

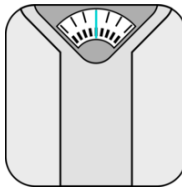
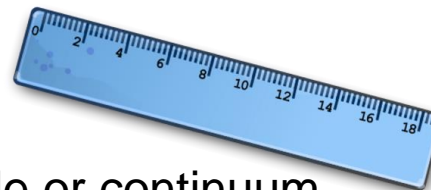


1 2 3
4 5 6
7 8 9



Discrete: the number of or proportion that fit into a category

Examples: Eye color, marital status, good/bad, boy girl, grade, objects that come in whole units(people, cars, animals, etc.)



Continuous: A number from a measurement scale or continuum

Examples: Weight, height, distance, money, time, temperature, length



Highlights: Video Segment: 1.7:Describing Data

3 Ways to Measure the Center of the Data

1. Mean(average): To find the mean of the values in a data set, simply add up all the numbers and divide by how many numbers you have. Sensitive to outliers.
2. Mode(most frequent value): The mode of a data set is the data value that occurs with the greatest frequency.
3. Median(middle point of the data): The median of a data set is the *middle data value* when the data are put into ascending order. Half of the data values lie below the median, and half lie above. If the sample size n is odd, then the median is the middle value. If the sample size n is even, then the median is the mean of the two middle data values. **Less sensitive to outliers than mean.**

3 Ways to Measure the Dispersion(or Spread) of the Data

1. Range: The range of a data set is the difference between the largest value and the smallest value in the data set:
2. Standard Deviation: may be interpreted as the typical difference between a data value and the sample mean for a given data set. Measures the spread of the data.
3. Variance: is approximately the mean of the squared deviations in the sample given by the formula, standard deviation squared

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Agenda

Topic	Time	Sunday Section	Wednesday Section
Introduction	5 min	6:30-6:35	9:00-9:05
Quiz 2 Prep	40 min	6:35-7:15	9:05-9:45
Highlights from Week 6 Video	40 min	7:15-7:55	9:45-10:25
Review of Upcoming Assignments and Open Question	5 min	7:55-8:00	10:25-10:30

Highlights: Video Segment 6.3: Causation Video

It might be useful to explain that "causes" is an asymmetric relation (X causes Y is different from Y causes X), whereas "is correlated with" is a symmetric relation.

For instance, homeless population and crime rate might be correlated, in that both tend to be high or low in the same locations. It is equally valid to say that homeless population is correlated with crime rate, or crime rate is correlated with homeless population. To say that crime causes homelessness, or homeless populations cause crime are different statements. And correlation does not imply that either is true. For instance, the underlying cause could be a 3rd variable such as drug abuse, or unemployment.

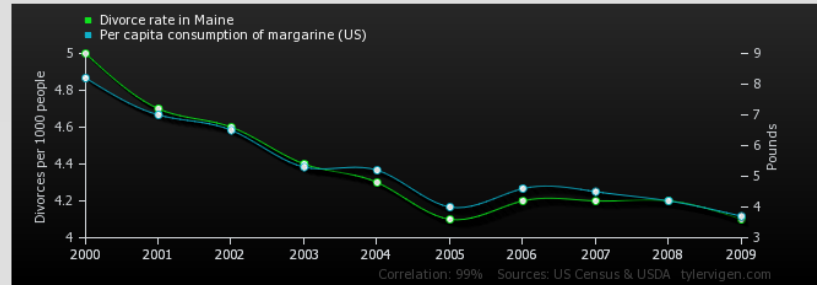
The mathematics of statistics is not good at identifying underlying causes, which requires some other form of judgement.

Causality is also a relationship between two things, but it is not mathematical, it is physical (or philosophical).

Something causes something else if there is a chain of events between the first thing and the second thing, each of which causes the next thing in the chain to happen. Causality involves time; the first thing happens, and then later the second thing happens as a result. We say the first thing is the cause, and the second thing is the effect. Note that unlike correlation, the relationship is unsymmetrical.

http://www.w-uh.com/posts/030302a_correlation_vs_ca.html

Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine Divorces per 1000 people (US Census)	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) Pounds (USDA)	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7
Correlation: 0.992558										

Permalink - Not interesting

<http://www.businessinsider.com/spurious-correlations-by-tyler-vigen-2014-5>


Correlation is the mathematical relationship between two things which are measured.

It is given as a value between 0 and 1. A correlation of 0 means the two things are unrelated; given the first value, there is no way to predict the second. A correlation of 1 means the two things are completely related, the first thing always predicts the second. As an example, let's say you measure the heights and weights of a group of people. These have a high correlation, somewhere around .8; height is a good predictor of weight, and vice-versa. Now say you took the same group and measured eye color. There is a low correlation between eye color and height, pretty close to 0. They are basically independent, knowing one doesn't tell you anything about the other.

Highlights: Video Segment 6.3:Causation Video

Acting on Correlation

Correlation may not indicate causation, but it can lead to action. Consider the marketing analyst that finds a strong correlation between site searches and revenue. To increase revenue, do we drive visitors to the search page on our site? Maybe, but what if visitors go to the site search page because they are frustrated and can't find the product they want to purchase? Do we want to increase frustration?

So what good is correlation if you are telling me I can't assume causation? Correlation gives you a strong starting point for further testing and optimization. Since we know there is a  relationship between site search and revenue, we can now look at why.

Through additional analysis, A/B tests, and customer surveys, we can better understand the relationship, context, and impact that our decisions will have on other factors.

So What?

Before diving into analysis, ask yourself "Why?" What is the business question and what value can be derived from my analysis? An important follow-up task is to consider the ability to take action and realize the value. As we get closer to understanding our target audience and what causes desired outcomes, remember the work is not finished until we act.

<http://blogs.adobe.com/digitalmarketing/analytics/hey-did-i-do-that-the-difference-between-correlation-and-causation/>

Highlights: Video Segment 6.4:Regression Intro

How Is Regression Useful?

- A way to see whether an input variable and output variable are related
- A statistical tool that can determine the strength of that relationship
 - Gives numeric value for strength of relationship where χ^2 test only tests for presence of relationship
- A model that predicts relationship between the two variables

Example of Use in Sales

A tool to increase sales efficiency:

- Predicting when a customer might cancel order/not purchase product
- Anticipating customers' growth (production volume)
- Understanding which customers to call and when

When you think of regression, think prediction. A regression uses the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable.

A simple linear regression uses only one independent variable, and it describes the relationship between the independent variable and dependent variable as a straight line.

https://www0.gsb.columbia.edu/premba/analytical/s7/s7_6.cfm

Regression Thought Process

1. **Practical:** Are these two variables practically linked?

2. **Graphical:** Plot the data!

- What does the scatter plot look like?

3. **Statistical:** Now perform statistical analysis.

Scatterplots

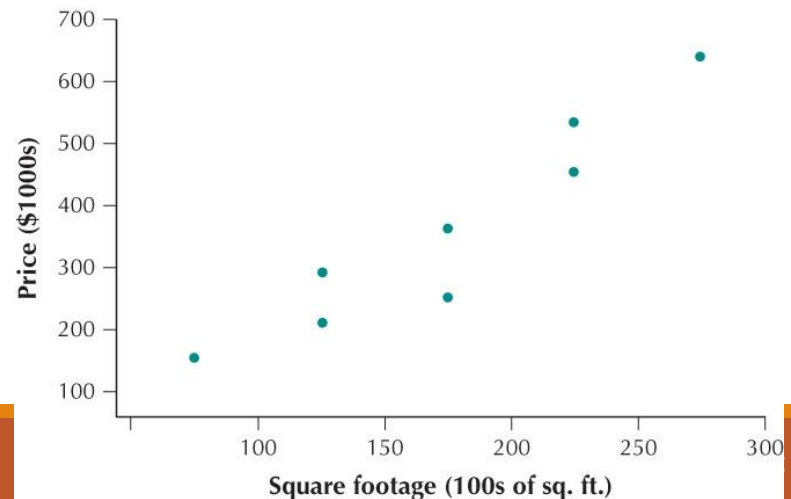
Whenever you are examining the relationship between two quantitative variables, your best bet is to start with a scatterplot. A **scatterplot** is used to summarize the relationship between two quantitative variables that have been measured on the same element.

A **scatterplot** is a graph of points (x,y) , each of which represents one observation from the data set.

One of the variables is measured along the horizontal axis and is called the *x variable*.

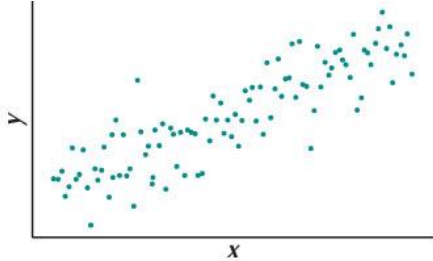
The other variable is measured along the vertical axis and is called the *y variable*.

Lot	x=square footage (100s of sq ft)	y=sales price (\$1000s)
Harding St	75	155
Newton Ave	125	210
Stacy Ct	125	290
Eastern Ave	175	360
Second St	175	250
Sunnybrook Rd	225	450
Ahlstrand Rd	225	530
Eastern Ave	275	635

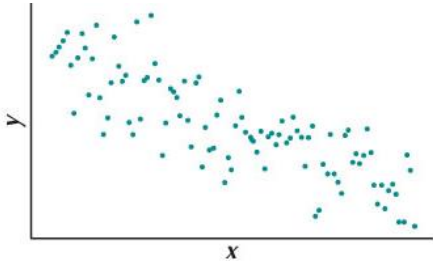


Scatterplots

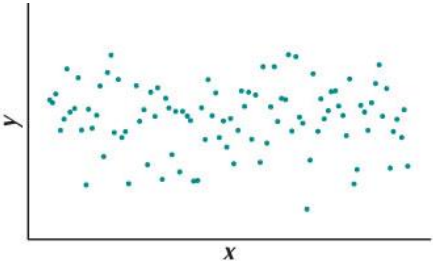
The relationship between two quantitative variables can take many different forms. Four of the most common are:



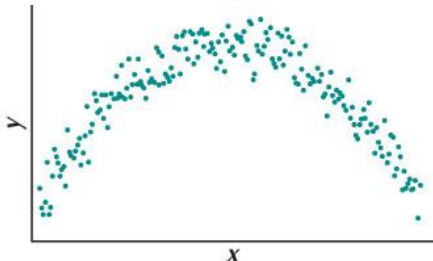
Positive linear relationship: As x increases, y also tends to increase.



Negative linear relationship: As x increases, y tends to decrease.



No apparent relationship: As x increases, y tends to remain unchanged.



Nonlinear relationship: The x and y variable are related, but not in a way that can be approximated using a straight line.

Simple Linear Regression: Equation

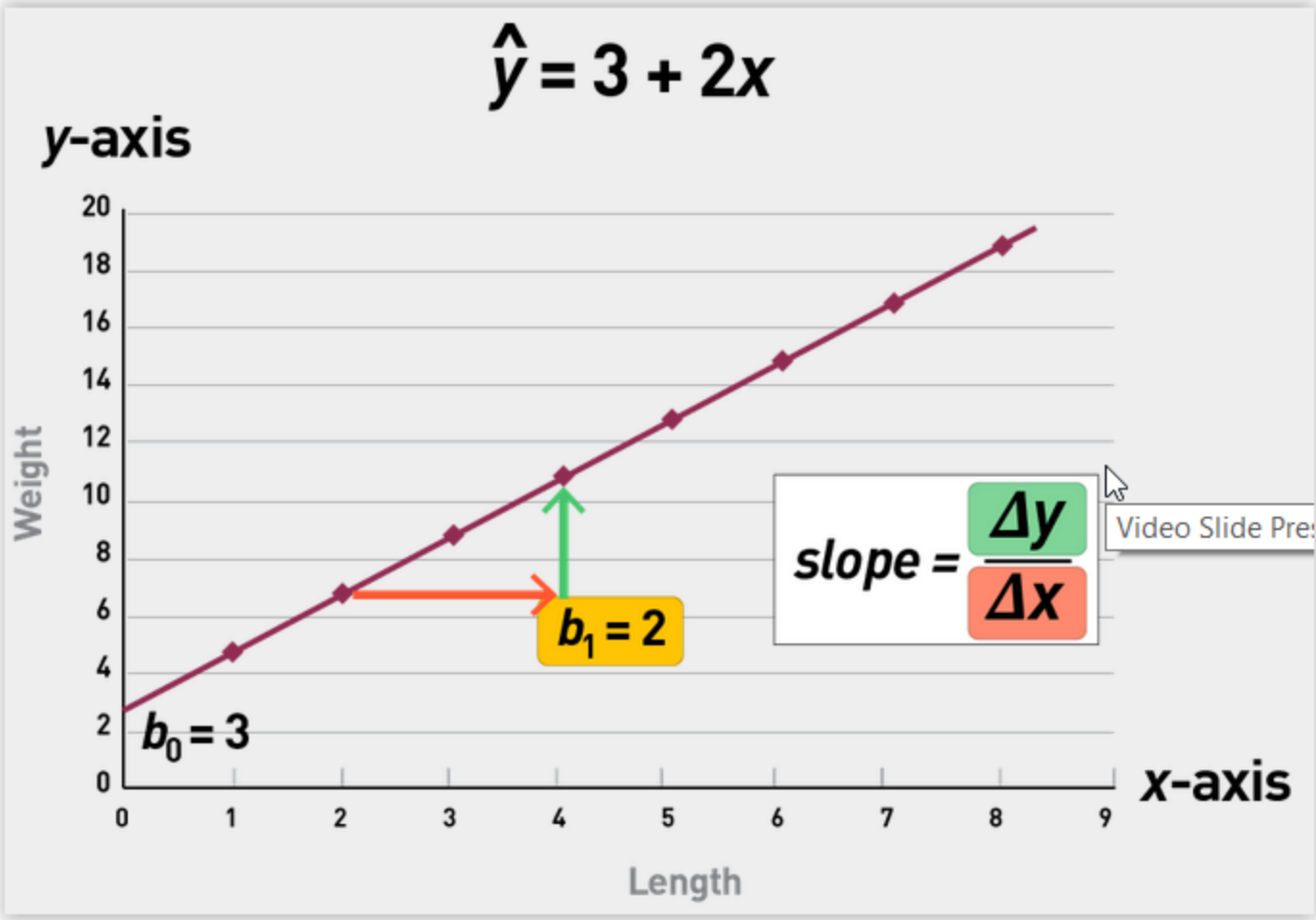
- Tries to create best-fitting line through plot
- Describes the relationship between two variables

$$\hat{y} = b_0 + b_1 x$$

Output variable
(predicted
response) y -intercept Slope Input variable

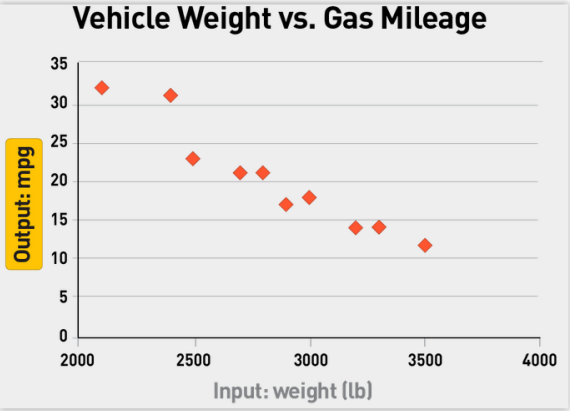
- **y -intercept:** where the line crosses the y -axis
- **Slope:** how slanted the line is

Highlights: Video Segment 6.4:Regression Intro



Highlights: Video Segment 6.5:Regression Example By Hand

Example problem:
How are the variables "vehicle weight" and "gas mileage" related?



Obs. (i)	x Weight (lb)	y Mileage (mpg)	$x_i y_i$	x_i^2
1	3,000	18	54,000	9,000,000
2	2,800	21	58,800	7,840,000
3	2,100	32	67,200	4,410,000
4	2,900	17	49,300	8,410,000
5	2,400	31	74,400	5,760,000
6	3,300	14	46,200	10,890,000
7	2,700	21	56,700	7,290,000
8	3,500	12	42,000	12,250,000
9	2,500	23	57,500	6,250,000
10	3,200	14	44,800	10,240,000
Totals	28,400	203	550,900	82,340,000
Averages	2,840	20.3		

Slope and Intercept Values

Obs. n = 10	x Weight (lb)	y Mileage (mpg)	$x_i y_i$	x_i^2
Totals	28,400	203	550,900	82,340,000
Averages	2,840	20.3		

• Slope:

$$\begin{aligned} b_1 &= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \\ &= \frac{550,900 - \frac{28,400(203)}{10}}{82,340,000 - \frac{(28,400)^2}{10}} \\ &= -0.0152 \end{aligned}$$

• Intercept:

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= 20.3 - (-0.0152)(2,840) \\ &= 63.5 \end{aligned}$$

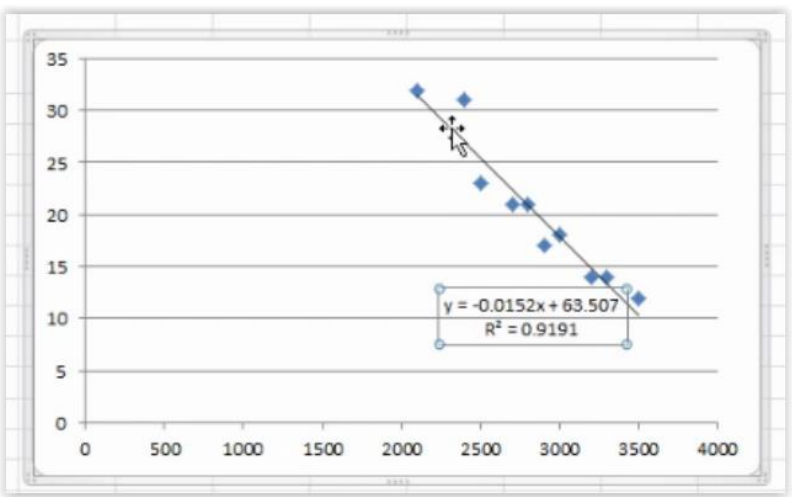
$$\hat{y} = 63.5 - 0.0152x$$

• Describes the relationship between vehicle weight and gas mileage

This formula can be used to predict gas mileage of any weight vehicle.

Highlights: Video Segment 6.6:Regression in Excel (p value and R square)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.95868061							
R Square	0.91906852							
Adjusted R Sq	0.90895209							
Standard Error	2.07132326							
Observations	10							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	389.777	389.777	90.84905	1.21E-05			
Residual	8	34.32304	4.29038					
Total	9	424.1						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	63.5071259	2.07132326	13.86565	7.08E-07	52.94522	74.06903	52.94522	74.06903
X Variable 1	-0.01521378	0.000896	-9.53148	1.21E-05	-0.0189	-0.01153	-0.01889	-0.01153



$$\hat{y} = 63.5 - 0.0152x$$

- Describes the relationship between vehicle weight and gas mileage

P-value = t distribution two-tailed probability, If one divides this by 2, it is the probability that the true value of the coefficient has the opposite sign to that found. You want this probability to be small in order to be sure that this variable really influences y, certainly less than 0.1 (10%). <http://people.clarkson.edu/~wwilcox/ES100/regrint.htm>

These are the probabilities that the coefficients are *not* statistically significant. http://www.udel.edu/johnmack/frec834/regression_intro.htm

Correlation Coefficient(r) is a measure of strength and direction and has values between -1 and 1.

Highlights: Video Segment 6.7:Correlation

Two Indices

1. Correlation coefficient (r)
2. Coefficient of determination (r^2)

Correlation Coefficient (r)

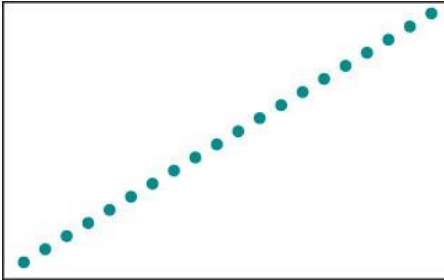
- $-1 < r$
- -1 = perfect negative correlation
- 1 = perfect positive correlation
- 0 = no relationship
- Rule of thumb: r value of $\sim \pm 0.7$ desired
- Indicates meaningful relationship

Scatterplots provide a visual description of the relationship between two quantitative variables. The *correlation coefficient* is a numerical measure for quantifying the linear relationship between two quantitative variables.

Properties of r

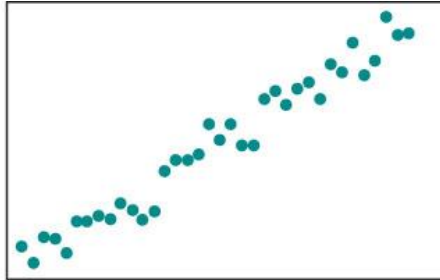
1. The correlation coefficient r is always $-1 \leq r \leq 1$.
2. When $r = +1$, a perfect positive relationship exists between x and y .
3. Values of r near $+1$ indicate a positive relationship between x and y .
 - The closer r gets to $+1$, the stronger the evidence for a positive relationship.
 - The variables are said to be **positively associated**.
 - As x increases, y tends to increase.
4. When $r = -1$, a perfect negative relationship exists between x and y .
5. Values of r near -1 indicate a negative relationship between x and y .
 - The closer r gets to -1 , the stronger the evidence for a negative relationship.
 - The variables are said to be **negatively associated**.
 - As x increases, y tends to decrease.
6. Values of r near 0 indicate there is no linear relationship between x and y .
 - The closer r gets to 0 , the weaker the evidence for a linear relationship.
 - The variables are **not linearly associated**.
 - A nonlinear relationship may exist between x and y .

Properties of r



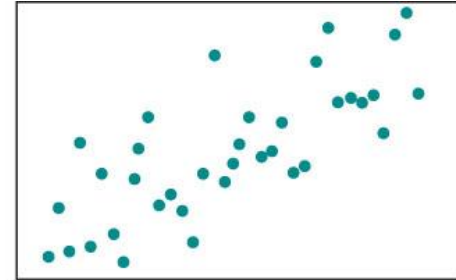
Perfect positive linear relationship, $r = 1$

(a)



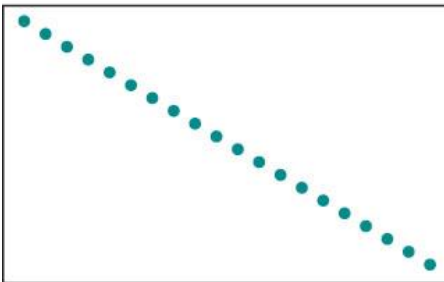
Strong positive linear relationship, $r = 0.9$

(b)



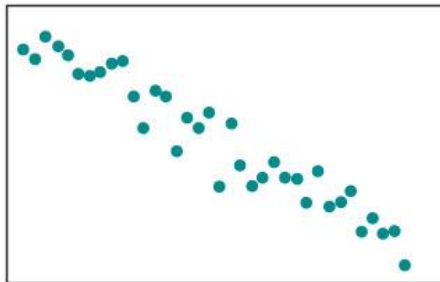
Moderate positive linear relationship, $r = 0.5$

(c)



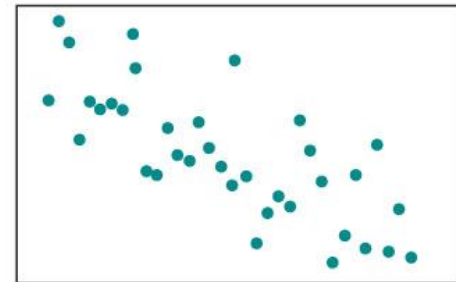
Perfect negative linear relationship, $r = -1$

(d)



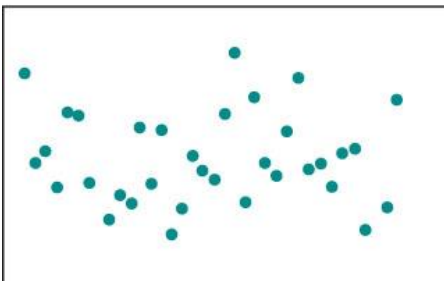
Strong negative linear relationship, $r = -0.9$

(e)



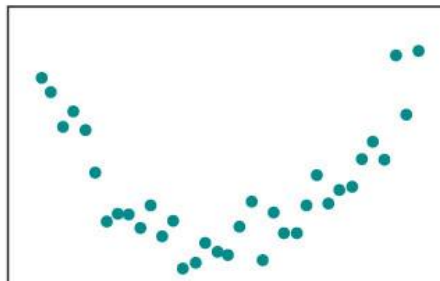
Moderate negative linear relationship, $r = -0.5$

(f)



No apparent linear relationship, $r = 0$

(g)



Nonlinear relationship but no linear relationship, $r = 0$

(h)

Highlights: Video Segment 6.7:Correlation

Two Indices

1. Correlation coefficient (r)
2. Coefficient of determination (r^2)

Coefficient of Determination (r^2)


- Correlation coefficient squared
- Measure of the percentage of variability in y that can be accounted for by x
 - Trying to find an input x that is influencing our output y
 - x will not explain all of y
 - Recall: There is variability in everything we do.
- Metric for whether input x is really contributing to output


Measures the goodness of fit of the regression equation to the data. We interpret r^2 as the proportion of the variability in y that is accounted for by the linear relationship between y and x . The values that r^2 can take are $0 \leq r^2 \leq 1$.

Highlights: Video Segment 6.7:Correlation

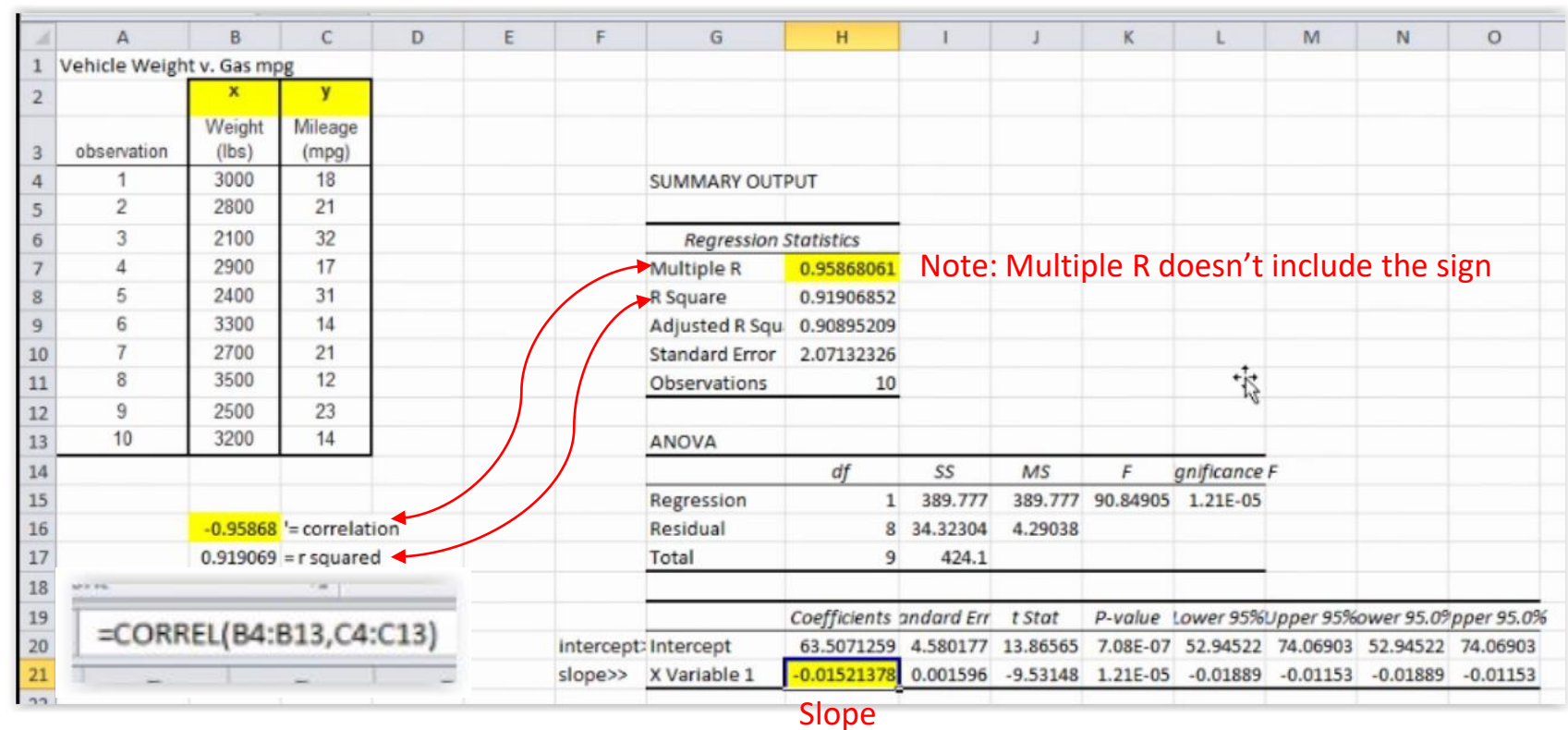
	A	B	C
1	SUMMARY OUTPUT		
2			
3	Regression Statistics		
4	Multiple R	0.958681	
5	R Square	0.919069	
6	Adjusted R Square	2.071323	
7	Observations	10	
8			

- Simple linear regression uses one x and one y .
- r^2 interpretation: 91% of variability in y is explained by x .

 r = Correlation Coefficient, what's the relationship between X and Y

 R Squared= Coefficient of Determination, how much variability in Y is explained by X

Highlights: Video Segment 6.8:Using Excel

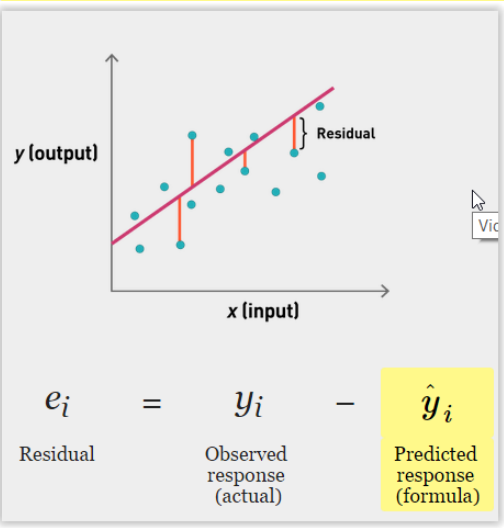


Multiple R = r, Correlation: get the sign from the X variable or separate correlation calc

Highlights: Video Segment 6.9:Residuals and Other Warnings

What Is a Residual?

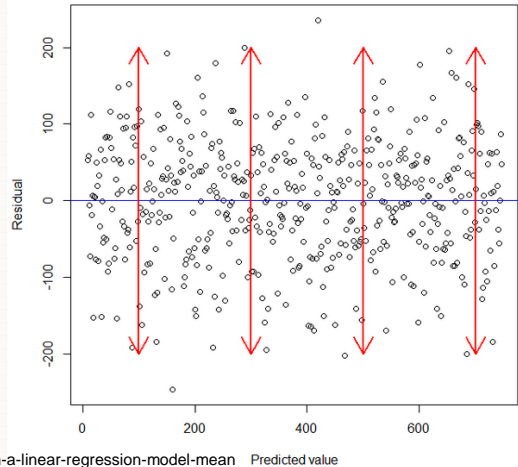
- Synonymous to error; should be random
- The distance between actual data point and the line determined by linear equation
- Determined by the difference between observed and predicted values of y
 - Ideally, points fall on regression line (i.e., perfect model)
 - Error would then be zero (rare).
- When plotted, a random series of points around a zero reference with no evidence of a pattern



Assumptions of Regression

1. Residuals are independent.
2. Residuals are normally distributed with a mean of zero.
 - The regression line will sometimes be high or low (i.e., over- or underpredicting).
3. There are equal variances (σ^2) of y .

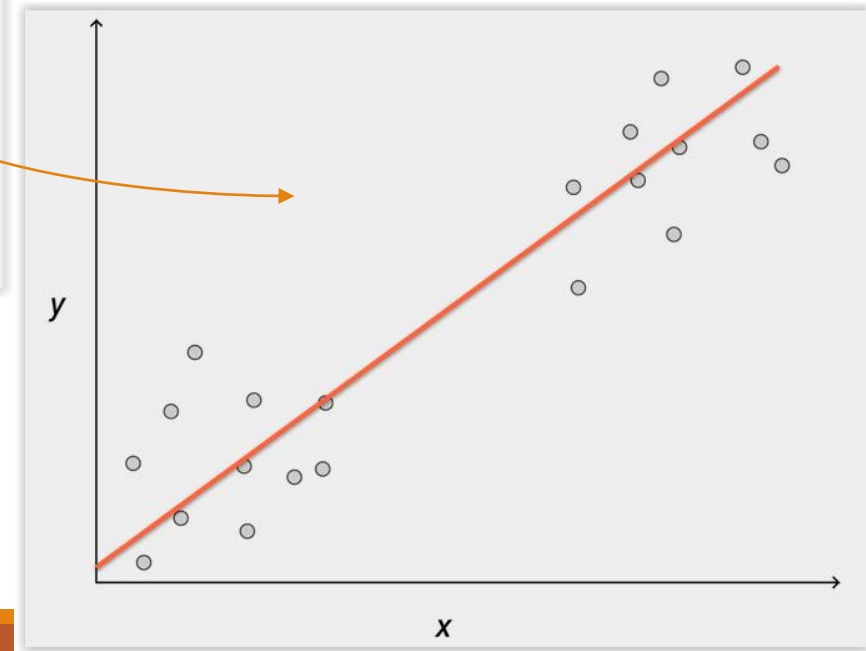
It means that when you plot the individual error against the predicted value, the variance of the error predicted value should be constant. See the red arrows in the picture below, the length of the red lines (a proxy of its variance) are the same.



Highlights: Video Segment 6.9:Residuals and Other Warnings

Other Points of Interest

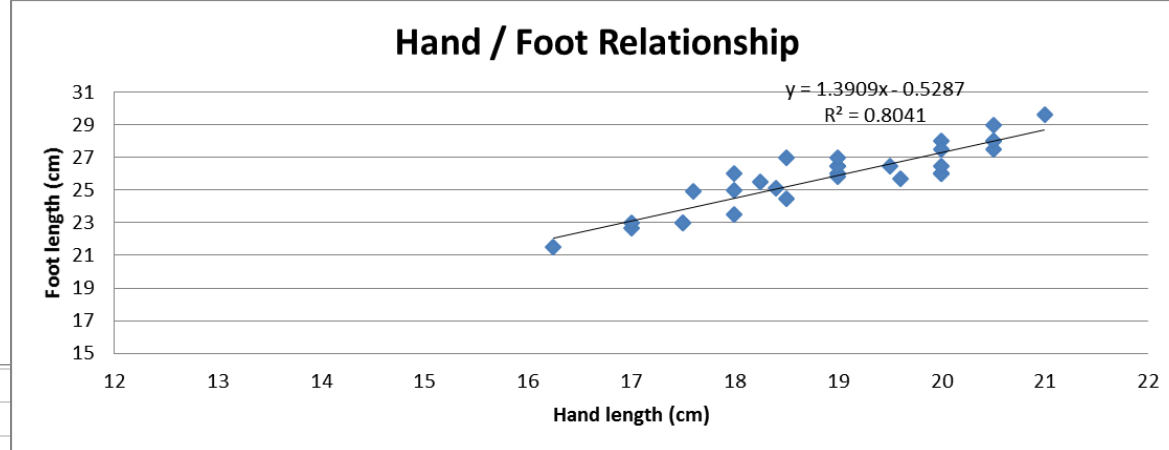
- Certain data is inappropriate for a regression analysis:
 - Residuals form a pattern.
 - Large outliers are present.
 - "Clumped" data appears linear.
- Avoid extrapolating outside data.
- Beware of lurking variables, or Simpson's paradox.
- A strong correlation does not mean causation.



Highlights: Video Segment 6.10:Residuals and Other Warnings Using Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Vehicle Weight v. Gas mpg														
		x	y											
observation		Weight (lbs)	Mileage (mpg)											
1		3000	18											
2		2800	21											
3		2100	32											
4		2900	17											
5		2400	31											
6		3300	14											
7		2700	21											
8		3500	12											
9		2500	23											
10		3200	14											
						SUMMARY OUTPUT								
						Regression Statistics								
						Multiple R	0.958680615							
						R Square	0.919068521							
						Adjusted R Sq	0.908952086							
						Standard Error	2.07132326							
						Observations	10							
						ANOVA								
							df	SS	MS	F	gnificance F			
						Regression	1	389.777	389.777	90.84905	1.21E-05			
						Residual	8	34.32304	4.29038					
						Total	9	424.1						
							Coefficients	Standard Error	t Stat	P-value				
						Intercept	63.50712589	4.580177	13.86565	7.0				
						X Variable 1	-0.015213777	0.001596	-9.53148	1.2				

Highlights: Video
Segment 6.11: Test
Your Knowledge:
Hand/Foot Exercise



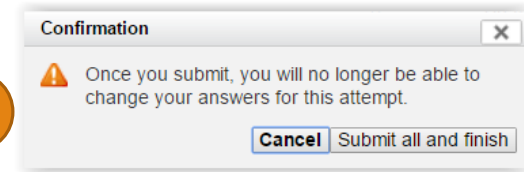
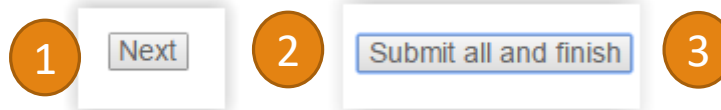
SUMMARY OUTPUT		12	13	14	15	16	17
		Hand length (cm)					
<i>Regression Statistics</i>							
Multiple R	0.89672317	r = correlation, what is the relationship.....positive sloping upward line					
R Square	0.804112444	Coefficient of determination....how much of the variability in foot measurements is explained by hand measurement					
Adjusted R Square	0.797582859						
Standard Error	0.850018382						
Observations	32						
ANOVA							
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>		
Regression	1	89	89	123.1490853	3.83668E-12		
Residual	30	22	0.7				
Total	31	111					
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95%</i> <i>Upper 95.0%</i>
Intercept	-0.528662862	2.4	-0	0.825767893	-5.390627992	4.333302269	-5 4.333302
X Variable 1	1.390895502	0.1	11	3.83668E-12	1.134923428	1.646867576	1.1 1.646868
Regression Equation	$y' = 1.39X - .52866$	Questions to answer:					
my actual hand size in inches	7.625	1) What is the length of your foot (without directly measuring it)?					
my actual hand size in cms	19.3675	Predict the length of your foot using the equation you just developed using the length of your hand (in centimeters) as the input.					
		2) What is your residual?					
Use formula to predict foot size $y' = 1.39X - .52866$		You will need to actually measure your foot to determine this.					
$y' = (1.39 \times 19.3675) - .52866$							
predicted foot size	26.392165000						
actual foot in cms	25.4						
residual	=Actual foot size - predicted foot size						
	=25.4-26.392165						
residual	-0.992165000	So the formula would have overestimated my foot size based on my hand size					

Agenda

Topic	Time	Sunday Section	Wednesday Section
Introduction	5 min	6:30-6:35	9:00-9:05
Quiz 2 Prep	40 min	6:35-7:15	9:05-9:45
Highlights from Week 6 Video	40 min	7:15-7:55	9:45-10:25
Review of Upcoming Assignments and Open Question	5 min	7:55-8:00	10:25-10:30

Review of Upcoming Assignments: Wednesday

1. Quiz #2 is due, Saturday, February 25th, midnight EST in the Learning Management System.
 - There are 5 calculation questions and 1, 10 part definitional question
 - DO NOT LEAVE ANY BLANK
 - You can not start and stop the Quiz.
 - There is a 2 hr. time limit. There is no timer, you must keep track of your own time.
 - Password for the Quiz is: **TestTime101**
 - At the end must click, **NEXT** at the bottom of the questions, then **click Submit all and Finish**, then click AGAIN **Submit all and Finish** in dialog box



2. Optional Learning: 7.9 Relate Regression to Your Project
3. HMWK #4 isn't due until 3/4 however it is a Learning Curve on Chapter 4: Correlation and Regression so you may want to do it after you are done with your Quiz #2 just to get it out of the way
4. Understanding Variation Book – start reading it before Live Class #8, Wednesday, 3/8
5. Projects - Analyze should be in full swing moving into Improve