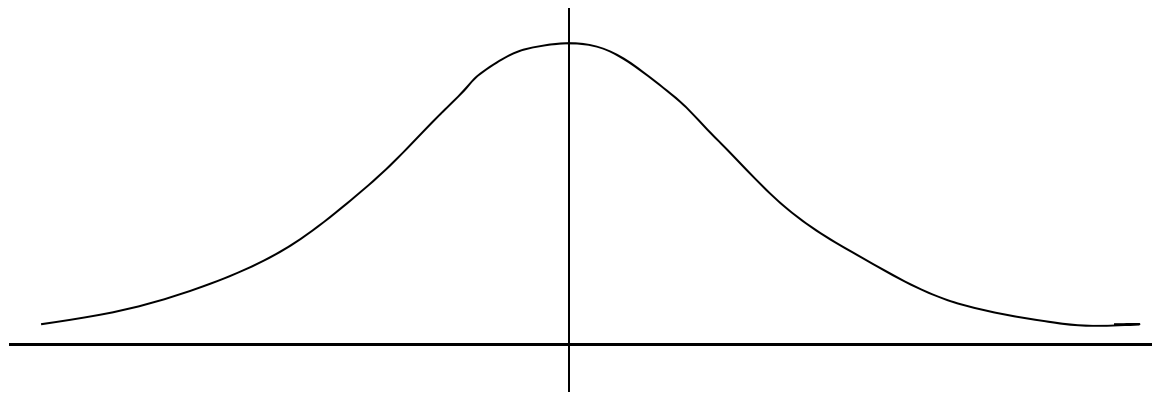


STAT 101

*An introduction to basic statistics in not nearly
enough time to do it justice*



Mark Berron, CQE, CSQE, SSBB
Perot System Government Services
Mark.Berron@psgs.com

Statistics (why we need it)

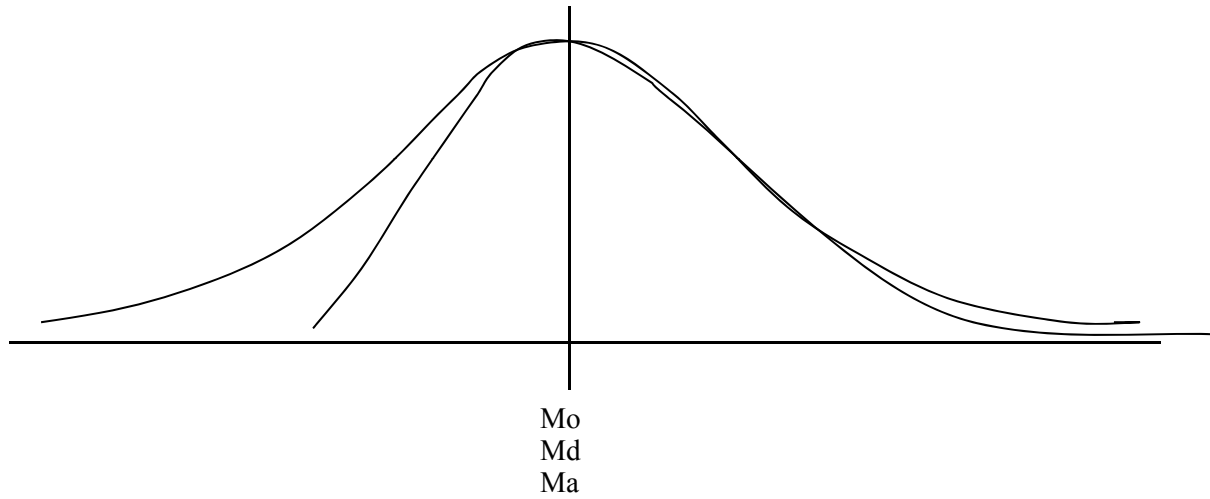
- **Statistics** is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data.
- Statistical methods can be used to summarize or describe a collection of data; this is called **descriptive statistics**.
- In addition, patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, and are then used to draw inferences about the process or population being studied; this is called **inferential statistics**.
- Statistics is basically the determination of the characteristics of a population based on the sampling of a smaller proportion.

Mean: What does it really mean??

- Mean (at least arithmetic mean) strictly means adding the numbers then dividing by how many numbers there are. In statistics, it is the sample mean that is usually referred.
- Average could be the mean, but it could also be the median if there are one or two values that distort the mean.
- Example: what does the average worker earn if these are the salaries: 20000, 20000, 20000, 21000, 21000, 21000 , 300000

Median: Why do we use it??

- Median is the middle number or average of the middle two numbers of a group.
- We can use the median too see what the middle number is. If it's close to the mean, then the distribution is probably normal. If a population is normal, we can do a lot with it statistically.



Mo = mode, Md = median, Ma = mean

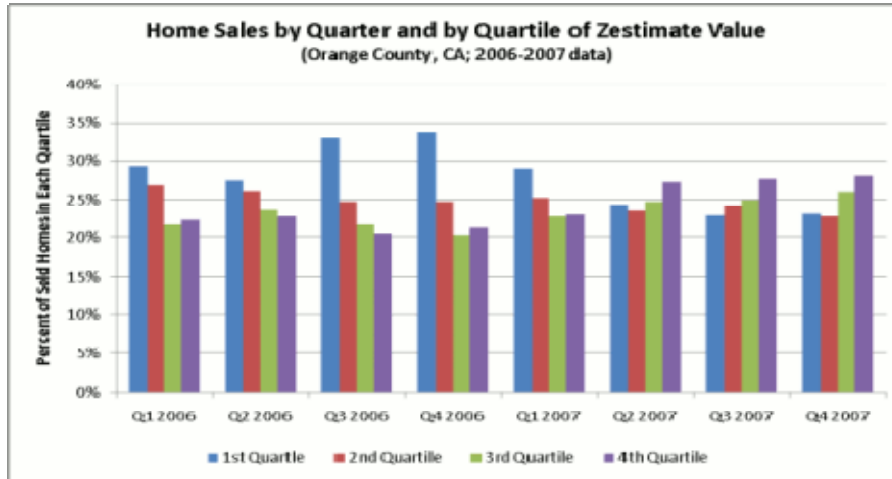
Median: Why do we need it ??

- To guard against outliers, and it doesn't let the far out values have higher weight.

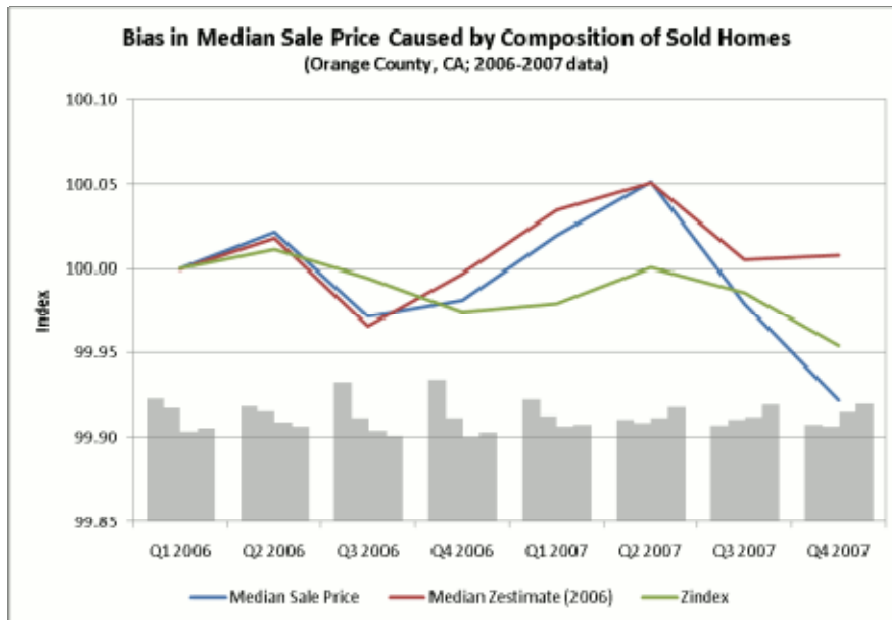
**Median and Average (mean) Sales Prices
Prices of New Homes Sold in the US**

Period	Median	Mean
Aug-04	218,100	272,200
Mar-07	262,600	329,400
Jul-07	246,200	298,200
Jun-08	231,700	296,800
Jul-08	234,900	299,100
Aug-08	221,900	263,900

What does this say??



- Note the distribution
- Note that median values are rising from Q12006 to Q2007.
- What happens to the distribution?



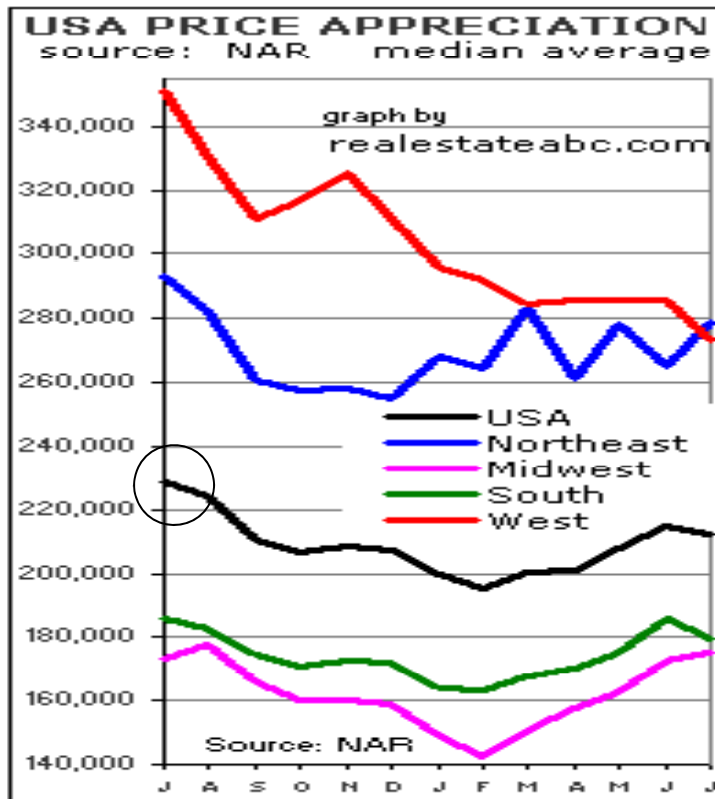
It starts to skew to the left in Q12007. What does that say about the average price?

Note the median price change is downward (blue line). The average price is going down faster.

Statistically, really ??

Real Estate Home Appreciation - Last 12 Months

Last Updated: 9/3/2008

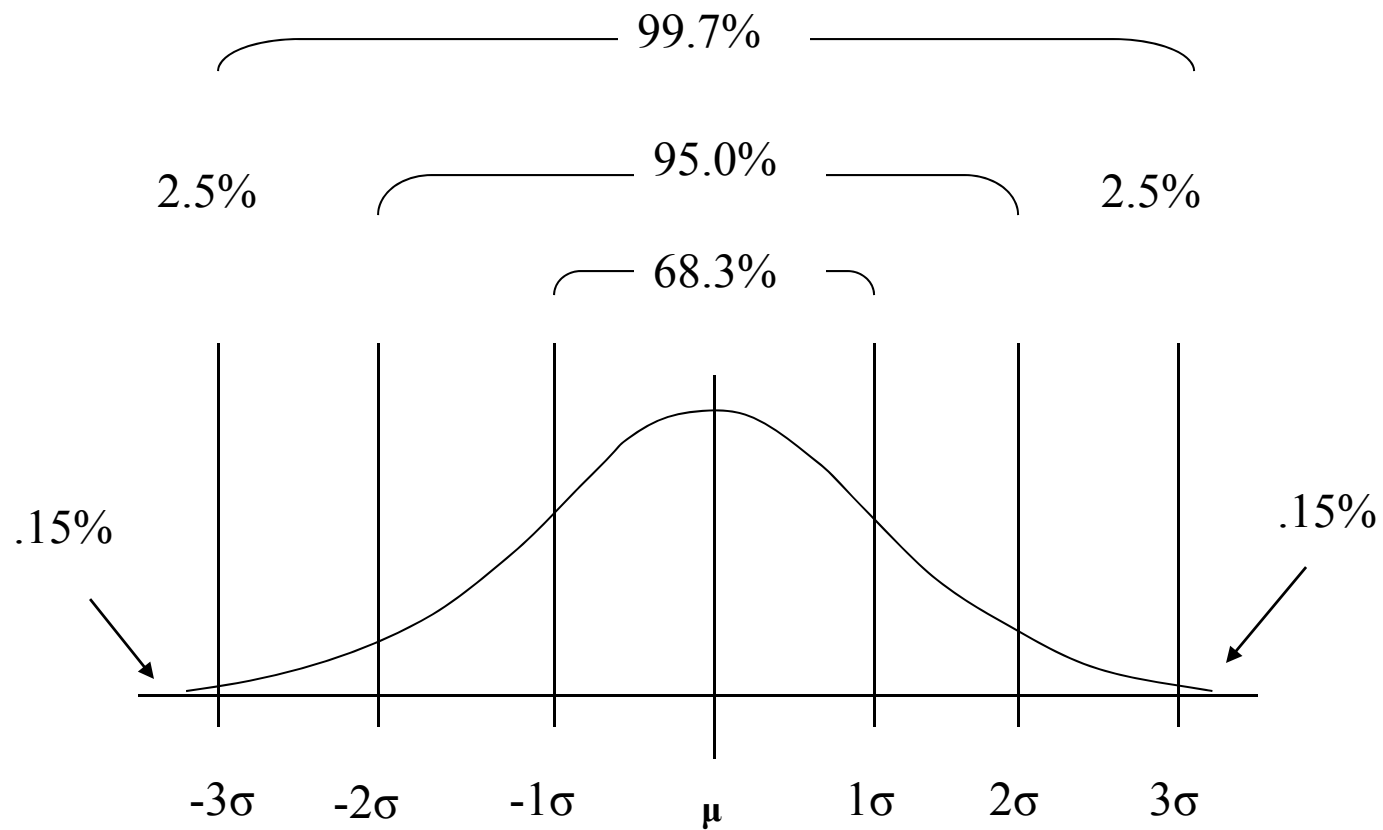


Home sales rise locally; average price falls by 22%.

Sales gains show positive trend for area market

<http://www.lansingstatejournal.com/apps/pbcs.dll/article?AID=20080817/NEWS03/808170562/1004/NEWS03>

NAR sales are based on mortgage company information, Census based on bank information. It confuses the averages consumer.



Normal distribution

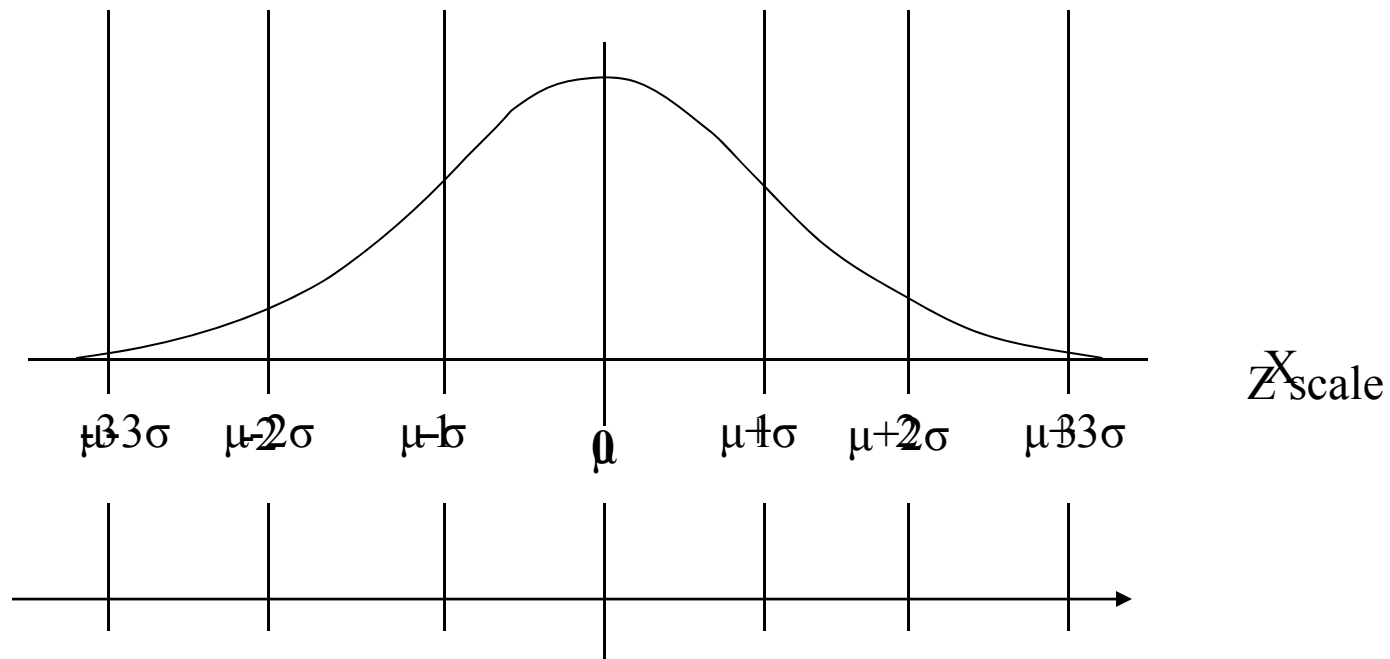
σ is the standard deviation for the population, which is the how far on average each point lies from the mean. Can be called the “error rate”.

Z, the Standard normal variate

- Z score is a useful tool that tells you where the set of observations are relative to the population within some probability

- $Z = (X - \mu) / \sigma$, where X is a value. When sampling, to compute Z, we

Divide by σ/\sqrt{n} , where n is the sample size.



Z scale example

- Say we want to know what % of a car dealership's inventory is below \$10m and given that the average inventory has \$12m and the standard deviation is \$1m.
- Get z. $Z = (10-12)/1 = -2$. So % below that z is what?

1092 Appendix

TABLE 1
Standard normal curve areas

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913

Use the z-table !

We are on the negative side, so take 1-.9772, or .0228, which is 2.28%.

OR...

The probability of having inventory less than \$10m is 2.28%.

P Values

P values give a quick idea of how significant a result might be.

If we are 99% confident that a result is less than a certain value, we say it has a p-value of .01, or 1-.99.

So, if we set the risk at 5%, or .05, and calculated a result that indicated the confidence level was 99%, or a pvalue of .01, then we know we are good. This leads us to hypothesis testing.

TABLE 2
Effect of SINGULAIR on Asthma-Related Outcome Measurements
(Combined Analyses - U.S. and Multinational Trials)

	SINGULAIR	Placebo
Asthma Attack* (% of patients)	11.6 [†]	18.4
Oral Corticosteroid Rescue (% of patients)	10.7 [†]	17.5
Discontinuation Due to Asthma (% of patients)	1.4 [‡]	4.0
Asthma Exacerbations** (% of days)	12.8 [†]	20.5
Asthma Control Days*** (% of days)	38.5 [†]	27.2
Physicians' Global Evaluation (score) [§]	1.77 [†]	2.43
Patients' Global Evaluation (score) ^{§§}	1.60 [†]	2.15
[†] p<0.001, compared with placebo		
[‡] p<0.01, compared with placebo		

* Asthma Attack defined as utilization of health-care resources such as

Hypothesis testing

Hypotheses can be for questions such as: Knowing whether 2 groups are really different. Knowing whether a group of values is less than or greater than a certain value.

For instance, say the desired level of temperature is 300 degrees in an oven, and we did 40 measurements with a normal distribution that showed an average or mean of 295 with a standard deviation (SD) of 45 degrees. You want to know if it really is less than the desired level with 95% confidence.

So this means you take the risk as 5%, and do your analysis.

Your hypothesis that the means are the same (null) is: $H_0 : \mu = 300$

Your alternate hypothesis is: $H_A : \mu < 300$

Now, we do our normal z calc:

$Z = (295-300)/(45/\sqrt{40}) = -.7027$; this converts to a probability of .2411. This translates into a p-value of .24 as well. This is well above the significance of .05, and therefore the null cannot be rejected. The means are the same.

If the SD was changed to 10, our zvalue would be -3.16. Translates into a pvalue of .0008, which is $< .05$, thus null is rejected and the means are different.

Risk, Error, and Confidence Intervals

- When we want to say that we are 95% confident the mean of a population is within certain values, we are saying there is a 5% risk we are wrong. This 5% risk is called Alpha risk. It is also called Type I error. To decrease the risk we increase the size of the sample, n. So to take risk into consideration, we must know the sample size.

- Confidence levels are established to give the public an opinion of how confident we are in the results. A summary might say, we are 95% confident that the mean of this population of say, squirrels in Baltimore County is between 1.3 million and 1.5 million, based on our random sample of the area.

- For Confidence Intervals then, the equation becomes:

$$CI = Y(\text{sample mean}) \pm Z_{1-\alpha/2} (\sigma/\sqrt{n})$$

Confidence Intervals Sample

What is the 95% confidence level for the mean weight of hereford cows based on the following:

The sample size is 50. The average weight is 659 lbs. with a Standard deviation of 45 lbs.

•For Confidence Intervals then, the equation becomes:

$$CI = Y(\text{sample mean}) \pm Z_{1-\alpha/2} (\sigma/\sqrt{n});$$

$$\text{So, CI 95\%} = 659 \pm Z_{1-0.025} * (45/\sqrt{50}) = (659 \pm 12.5)$$

$$(646.5, 671.5)$$

TABLE 1
Standard normal curve areas

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854

Summary:

Therefore, we are 95% confident that the true mean weight of the hereford cow population is between 646.5 and 671.5.

Multivariate analysis

Multivariate analysis basically gets a lot of data for various variables, and calculates whether they are significant or not to what the data model is trying to explain.

When a model is completed, using programs such as SAS or Minitab, and techniques such as Regression Analysis, Canonical Discrimination or Linear Discrimination, the most significant factors can be determined and a good model be made.

Models may include:

How do you explain the results of a clinical study among many different variables?
Was the tested variable really significant?

If you have 4 different species of cows, which ones give the most milk and why if you vary the hay, the temperature, the elevation?

By taking measurements of birds in different parts of its wings, neck circumference, can you tell if there are distinct species?

By taking measurements of baseball players from different eras, can you tell how a player will play today?

By looking at demographics of voting patterns in the past, can you tell what will happen in the election of 2008?

A look at Minitab for Regression Analysis (it won't hurt)

Regression Analysis: Share versus x1, x3, x2x3, x3x4

The regression equation is

$$\text{Share} = 3.39 - 0.416 x_1 + 0.394 x_3 - 0.000267 x_2x_3 + 0.200 x_3x_4$$

Predictor	Coef	SE Coef	T	P
Constant	3.3873	0.4003	8.46	0.000
x1	-0.4159	0.1713	-2.43	0.021
x3	0.3939	0.1043	3.78	0.001
x2x3	-0.0002671	0.0002564	-1.04	0.305
x3x4	0.20027	0.06980	2.87	0.007

Which does not fit here???

Regression Analysis: Share versus x1, x3, x3x4

The regression equation is

$$\text{Share} = 3.20 - 0.334 x_1 + 0.308 x_3 + 0.176 x_3x_4$$

Predictor	Coef	SE Coef	T	P
Constant	3.1959	0.3562	8.97	0.000
x1	-0.3336	0.1523	-2.19	0.036
x3	0.30808	0.06412	4.80	0.000
x3x4	0.17623	0.06597	2.67	0.012