

3190 Week 4

Sampling & Normal probability

Normality

- A random sample from a population that is normally distributed will be normally distributed
- *Asymmetry matters only for small samples from non-normal populations*

Assuming normality

- We would like to be able to assume normality
 - Then we can use parametric statistics, which are more powerful
 - For example, more likely to determine a difference or see a relationship
 - More powerful because we can use the normal probability distribution to make predictions
- *If our sample is random, we can assume normality at samples $n \geq 30$, why?*

Sampling

- Has to do with the nature of sampling and probability
- Before we learn about the magic number, $n \geq 30...$
- Let's review basic probability & sampling

Why is sampling important?

- When we need data to answer a question we have three options
 - Censuses
 - Experiments
 - Samples
- As you know statistical analyses use samples
 - It is critical that those samples represent populations well... called **representative sampling**

Classic example of poor sampling

- 1936 presidential election
- Republican Alfred Landon predicted to win in a landslide over Franklin D. Roosevelt by *Literary Digest*
 - Based on a poll (a sample of the American population)
- FDR won in a landslide, what happened?
 - Two **biases** in the sample
 - 1) Sample obtained among people who owned a car or telephone = wealthy in 1936 (tended to vote Republican)
 - 2) Only 25% polled responded; there was a non-response bias
 - Those who did not respond tended to vote for FDR.

Also important

- All inferential tests rely on the assumption that samples are representative
- Especially so for parametric tests... why?
 - *Because we are assuming **normality**, a characteristic of the population*
- Larger samples tend to be more representative, why?
 - *Because smaller samples **do not capture enough variability** to be representative*

Remember...

- The central goal of inferential statistics is...
- To draw conclusions about a population based on a sample
- Before we discuss inferential tests, we must ensure that we know how to produce representative samples

Probability Sampling: *A general* category

- The easiest way to ensure representation is to choose one of several “probability” or “random” sampling techniques
- In all probability sampling techniques a **random device** is used to decide which members of a population are included
 - Replaces human judgment (subjective choice).

Essential Concepts

- **Target population** = the complete set of individuals that a sample will represent
- **Target area** = a geographic twist, the entire region of set of locations that a sample will represent
- **A sampling frame** = the operational set that contains the entire set of cases from which a subset of cases will be drawn
 - = *the practical population*, can be locations (area) or individuals (population)
 - It's the entire set of cases (whatever they might be) that you will draw a sample from

Simple Random Sampling

- A probability sampling technique in which each case (individual) in the sampling frame has an equal chance of being selected
- Each case in the sampling frame must be identifiable to facilitate its random selection, usually by a number (e.g., Case # 202).
- We use a random number table to choose simple random samples

Simple Random Sampling, example

- Dr. Oppong knows that student evaluations can be misleading in terms of instructor performance
- Of the 728 students who have taken World Regional Geography during the last few years, he wants to conduct interviews
- He can interview only a small number of students
 - He settles on 15 randomly selected students

His sampling strategy

- He sets up a sampling frame numbering each student from 001 to 728.
- He could just pick the first 15 or the last 15 or students he knows, but he wants to cover multiple semesters and to be unbiased
- So he decides to use a random number table to produce a simple random sample

Picking the first number

- Dr. Oppong closes his eyes...
- And puts his finger on a number on the page...
- Then he uses the table to help him pick the fifteen students he wishes to interview
- Here's how...

TABLE B**Table of Random Numbers**

31871	60770	59235	41702	89372	28600	30013	18266	65044	61045
87134	32839	17850	37359	27221	92409	94778	17902	09467	86757
06728	16314	81076	42172	46446	09226	96262	77674	70205	98137
95646	67486	05167	07819	79918	83949	45605	18915	79458	54009
44085	87246	47378	98338	40368	02240	72593	52823	79002	88190
83967	84810	51612	81501	10440	48553	67919	73678	83149	47096
49990	02051	64575	70323	07863	59220	01746	94213	82977	42384
65332	16488	04433	37990	93517	18395	72848	97025	38894	46611
42309	04063	55291	72165	96921	53350	34173	39908	11634	87145
84715	41808	12085	72525	91171	09779	07223	75577	20934	92047
63919	83977	72416	55450	47642	01013	17560	54189	73523	33681
97595	78300	93502	25847	19520	16896	69282	16917	04194	25797
17116	42649	89252	61052	78332	15102	47707	28369	60400	15908
34037	84573	49914	59688	18584	53498	94905	14914	23261	58133
08813	14453	70437	49093	69880	99944	40482	04254	62842	68089
67115	41050	65453	04510	35518	88843	15801	86163	49913	46849
14596	62802	33009	74095	34549	76634	64270	67491	83941	55154
70258	26948	60863	47666	58512	91404	97357	85710	03414	56591
83369	81179	32429	34781	00006	65951	40254	71102	60416	43296
83811	49358	75171	34768	70070	76550	14252	97378	79500	97123
14924	71607	74638	01939	77044	18277	68229	09310	63258	85064
60102	56587	29842	12031	00794	90638	21862	72154	19880	80895
33393	30109	42005	47977	26453	15333	45390	89862	70351	36953
92592	78232	19328	29645	69836	91169	95180	15046	45679	94500
27421	73356	53897	26916	52015	26854	42833	64257	49423	39440
26528	22550	36692	25262	61419	53986	73898	80237	71387	32532
07664	10752	95021	17030	76784	86861	12780	44379	31261	18424
37954	72029	29624	09119	13444	22645	78345	79876	37582	75549
66495	11333	81101	69328	84838	76395	35997	07259	66254	47451
72506	28524	39595	49356	92733	42951	47774	75462	64409	69116
09713	70270	28077	15634	36525	91204	48443	50561	92394	60636
51852	70782	93498	44669	79647	06321	04020	00111	24737	05521
31460	22222	18801	00675	57562	97923	45974	75158	94918	40144
14328	05024	04333	04135	53143	79207	85863	04962	89549	63308
84002	98073	52998	05749	45538	26164	68672	97486	32341	99419
89541	28345	22887	79269	55620	68269	88765	72464	11586	52211
50502	39890	81465	00449	09931	12667	30278	63963	84192	25266
30862	61996	73216	12554	01200	63234	41277	20477	71899	05347
36735	58841	35287	51112	47322	81354	51080	72771	53653	42108
11561	81204	68175	93037	47967	74085	05905	86471	47671	18456

The number he picked

I zoom in on this section in the next slide so you can see it

TABLE B

Table of Random Numbers

31871	60770	59235
87134	32839	17850
06728	16314	81076
95646	67486	05167
<u>44085</u>	87246	47378
83967	84810	51612
49998	02051	64575
65332	16488	04433
42309	04063	55291
84715	41808	12085
63919	83977	72416
97595	78300	93502
17116	42649	89252
34037	84573	49914
08813	14453	70437
67115	41050	65453
14596	62802	33009
70258	26948	60863
83369	81179	32429
83811	49358	75171

Begin with the starting point

Your frame is from 001 to 728, so you need three digit numbers

Cross out the two numbers on the right side of 95646; this leaves 956

- 956 is out of your frame

Move down one number, 44085; cross out 8 & 5 leaving 440

- 440 is in your frame, it is the first of your 15 cases, 14 left

Move down one number, 83967; 839 is not in your frame, move down one more, 499 is so pick it, **and so forth...**

The Sample

- Dr. Oppong would interview students # 440, 499, 653, 423, 639, 171, 340, 088, 671, 145, 702, 149, 601, 333, 274

The result is a group that is randomly selected and thus more likely to be representative

That is, there is no *biasing choice mechanism* in the sampling.

Uneven coverage & \$\$ costs: problems?

- Because simple random sampling is completely random, there is no guarantee of even coverage of the sampling frame
- Additionally, it can be costly in geography to travel to sample
- There are a variety of sampling strategies to deal with these problems

Systematic sampling – guarantees even coverage

Stratified sampling – very useful for populations/ areas with different subsets to them

Cluster sampling minimizes costs and targets efforts (very important in geography)

Multistage sampling may combine advantages of approaches

Systematic Sampling

- Sampling that starts with ordering the **case labels** from lowest to highest then picks the first case randomly and selects at an equal interval for the rest of the cases
- For Dr. Oponng, number each student and order from 001 to 728
- Pick the first case and following cases ...
 - Determine the interval size = K
 - Pick the first case randomly from the first interval

Determining the Interval

- Calculate the interval (K) based on the desired sample size
- We desired a sample of 15; to determine the interval take

$$728/15 = K$$

$$K = 48.5 \text{ (round to 48)}$$

Always round down in SIS

- Pick the first case from 001 to 048 randomly using the random number table
- Then add 48 to that first case to get the next one, and so forth

TABLE B**Table of Random Numbers**

31871	60770	59235
87134	32839	17850
06728	16314	81076
95646	67486	05167
44085	87246	47378
83967	84810	51612
49990	02051	64575
65332	16488	04433
42309	04063	55291
84715	41808	12085
63919	83977	72416
97595	78300	93502
17116	<u>42649</u>	89252
34037	84573	49914
08813	14453	70437
67115	41050	65453
14596	62802	33009
70258	26948	60863
83369	81179	32429
83811	49358	75171

This time we are picking a random number from 01 to 48 (the first interval).

Close your eyes and put your finger on the table...

Let's say we land on 83; it is not in the interval...

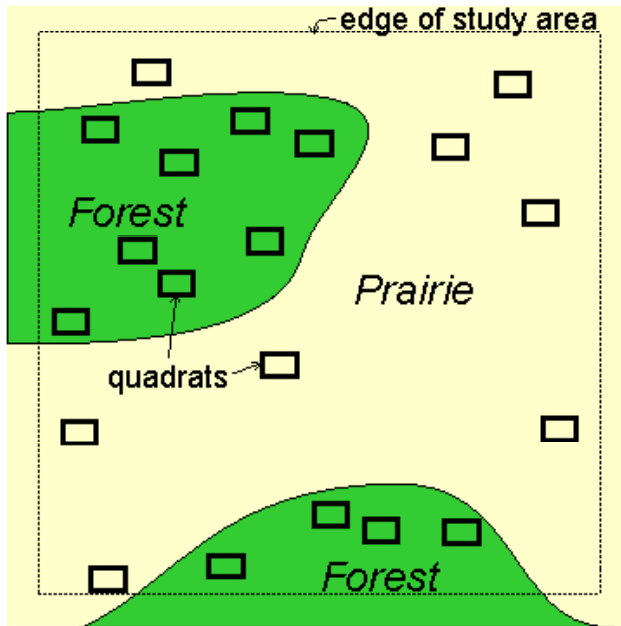
But move down to the first number that is

It is 42, which is your first case

Add 48 to 42, and 90 is your next case

42, 90, 138, 186, 234, 282, 330, 378, 426, 474, 522, 570, 618, 666, 714

Stratified Random Sampling



oregonstate.edu/instruct/bot440/wilsomar/Content/Assets/StRS.gif

- A method of sampling that takes into account known differences in the underlying population
- Here the target population is separated into several groups (strata) to reflect that underlying structure
 - Called “target subdivision”
 - A random device is then used to sample strata
- This sample is stratified into forest and prairie

Two kinds

Proportional stratified random

- The same proportion of area or population is sampled in each stratum
- Let's say I wanted to sample plots to determine community vegetation in the prairie and forest areas
- I need to find out equally about both strata

Disproportional stratified random

- A higher proportion of a stratum is sampled than for other strata
- Let's say I wanted to learn about the abundance of a bird species that occurs most often in the forest, but less so in prairie
- I need to sample both areas, but forest more-so

Other examples

Proportional

- Voting preferences & residence types, 10% sample
- I want to make sure I cover all types of residences and sample each randomly
- Stratify by type: apartments, houses, condominiums, mobile home, etc...
- Take a 10% sample from each

Disproportional

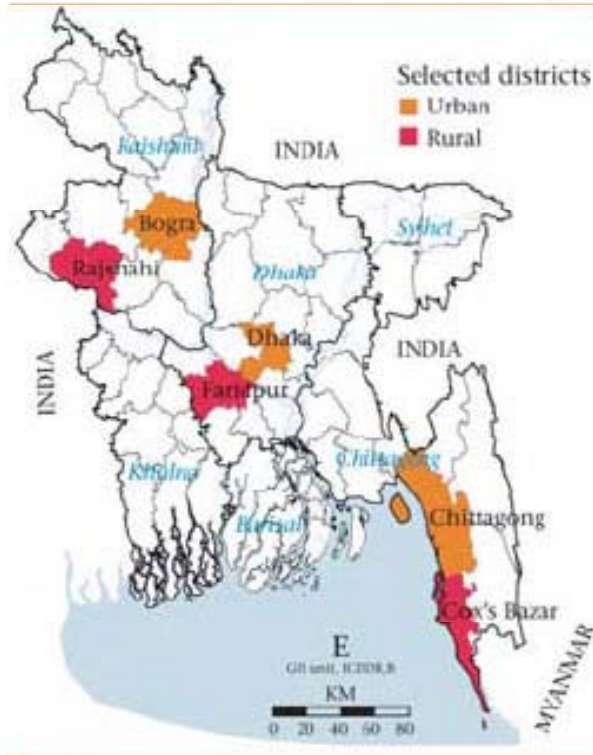
- Let's say legislation to be voted upon is most important to house owners
- I would still want to sample each stratum (residence type)
- But I might take a 20% sample from homeowners and less (e.g., 5% from others)

Stratified random sampling

- You decide on the appropriate subdivision based on the questions you ask
- The key is to sample **within each stratum** randomly
 - Can be done with simple random sampling
 - Or with systematic sampling

Cluster Sampling

Figure 1: Study areas (selected districts)



- A method of sampling in which cases are selected from **groups** within the sampling frame
- In this study of HIV transmission in Bangladesh, researchers studied rural and urban areas
 - Within those areas, simple random sampling would have been inefficient
 - They chose clusters (neighborhoods, villages) and studied 30 clusters in each area

To cluster sample

- Divide population into groups (clusters)
- Randomly select a subset of those clusters
- Collect data within selected clusters
 - Either **census** within the cluster
 - Or **randomly sample** within the cluster (2 stage)

Cluster sampling: another example

- Let's say we want to sample parasites in horses in North Texas to determine risk for a new rancher
- We could randomly select USGS sections then go look for horses in the sections we select
 - Inefficient, why?
- Or, we could pick multiple areas (clusters) where horses are ranched, randomly select a subset of clusters and then study ranches within each cluster
 - Efficient, why?

Cluster sampling

- Very efficient in geography where **sampling often requires travel**
- For example, suppose the 728 students in Dr. Oppong's sampling frame were all over the world after they graduated...
 - Wouldn't it be most efficient to randomly select a subset of large cities and then randomly sample alumni in those areas?
- Depending on \$\$ & time you may **sample every case within a cluster or randomly sample within each cluster**

Multistage sampling

- Complex sampling designs that combine one or more of the traditional approaches
- Cluster sampling can be multistage if you sample within clusters
 - If you census within cluster then it is not
- Example, you might **stratify** an area into subsections, **randomly select clusters** within each stratum, and then **systematic sample within each cluster**

Normal Probability

Inferential Statistics

- Rely on probability theory
- Up until now, all descriptive
 - But we would like methods with which to draw inferences about a population using a **sample**
 - Because we use part of the population to draw inferences about the whole population there is always **uncertainty** in the correctness of our conclusions = **error**

Probability Theory

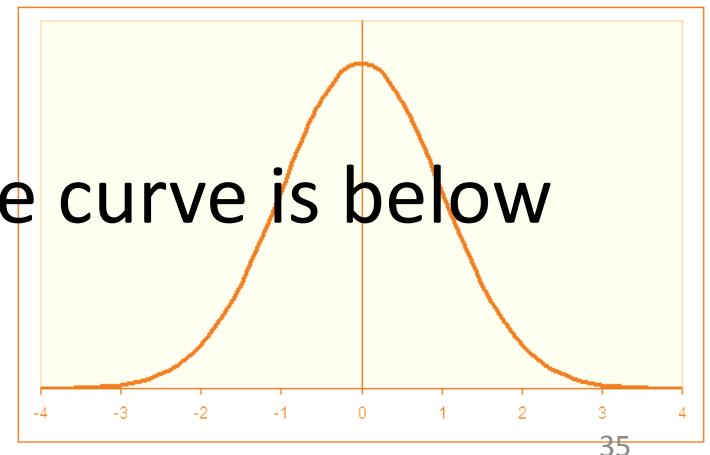
- Is the science of uncertainty
- “Enables us to evaluate and control the likelihood that a statistical inference is correct” (Weiss 2002:146).
- Probability = the chance that any particular outcome for an event will take place

Properties of Probability

- The probability of an outcome is always between 0 and 1
- The probability of an outcome that cannot occur is always 0, an **impossible** outcome
- The probability of an outcome that must occur is 1, a **certain** outcome

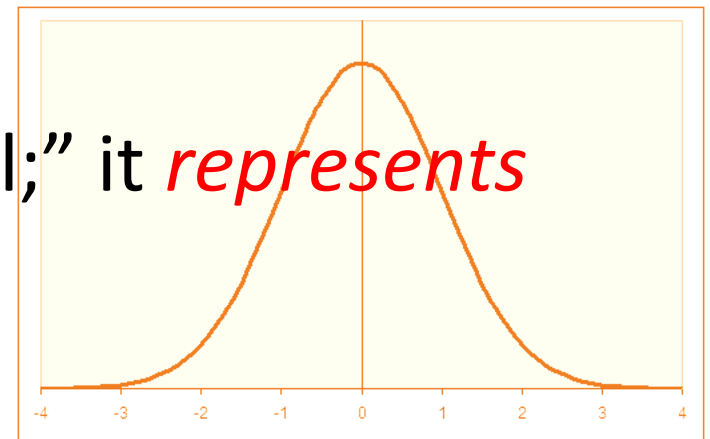
Area & the Normal Curve

- The total area under the curve = 1
- So if we asked the question, what is the probability of encountering a case at the mean or less?
 - It would be 0.5 because the mean is the middle of the curve
- That is, half of the area of the curve is below the mean, to the left.



What is the “normal” curve

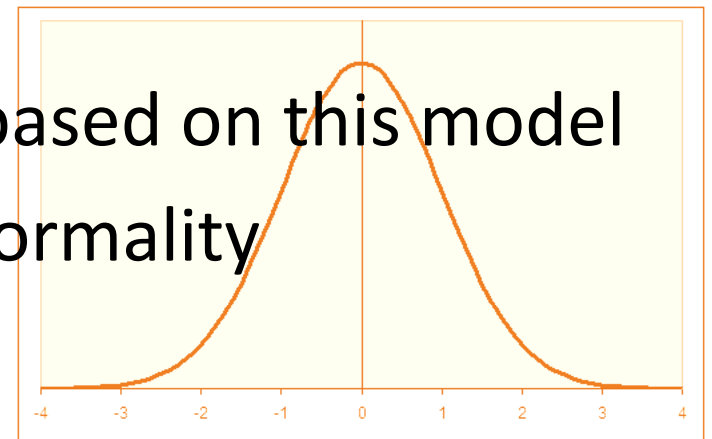
- It is a model of the perfect symmetrical distribution
- It was derived mathematically
- Its purpose is to serve as an *ideal example* of the a data distribution we tend to see often
 - Symmetrical
 - Unimodal
- The normal curve is not “real;” it *represents* reality



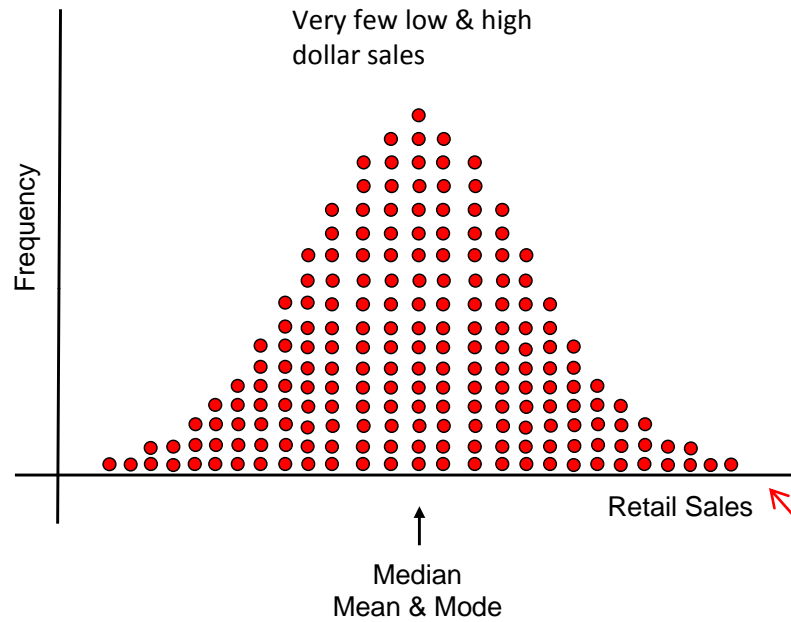
What does it mean to assume normality?

- The normal curve is an ideal (model)
- Area under the curve is used to make predictions
- In order to make predictions with our real data **we must assume** that our data are normally distributed

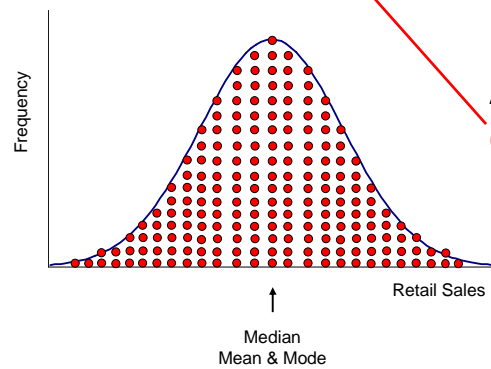
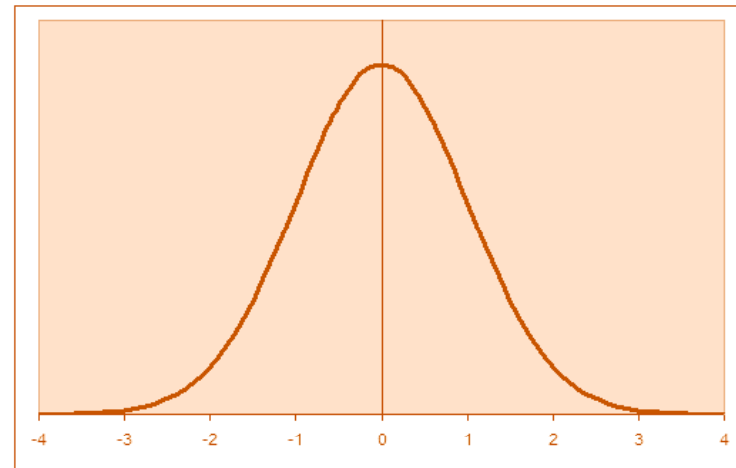
- Variance, Standard Deviation based on this model
- Parametric statistics assume normality



Perfectly symmetrical, real distribution
A normally distributed distribution



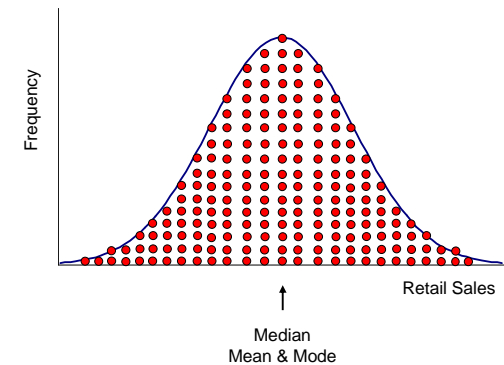
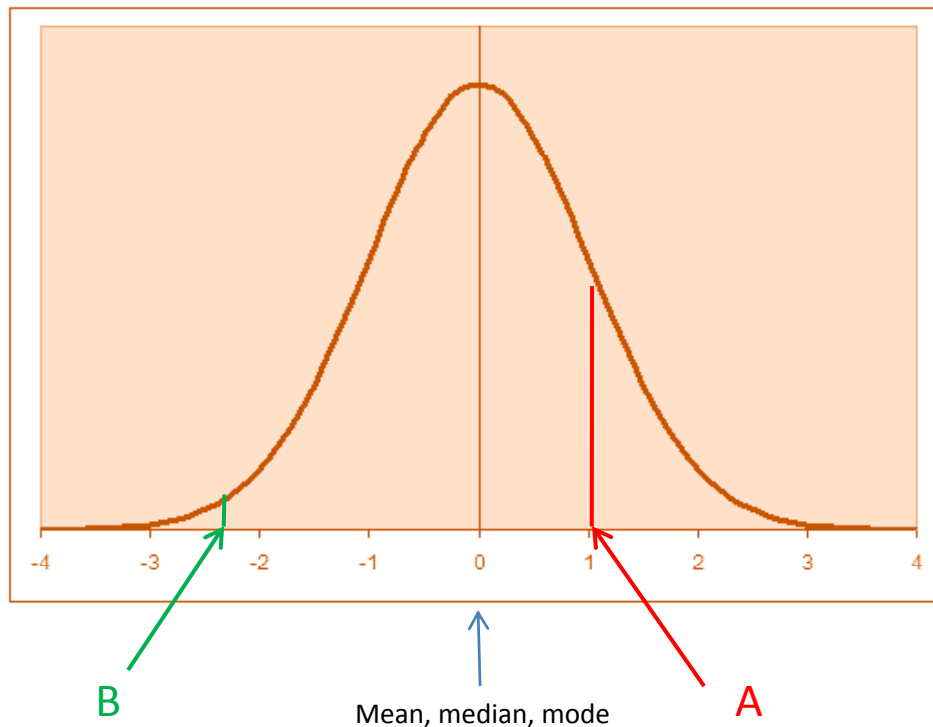
A model of a perfectly symmetrical distribution
A "normal curve"
The basis of parametric description & inference



A normally distributed real **data** distribution with a superimposed normal curve

Probability

- We often speak of probability when using the normal curve
- Area under a portion of the curve is the probability of encountering a particular score
- There is a higher probability of encountering a score at **A** than **B**
 - It is at a part of the curve with more area under it

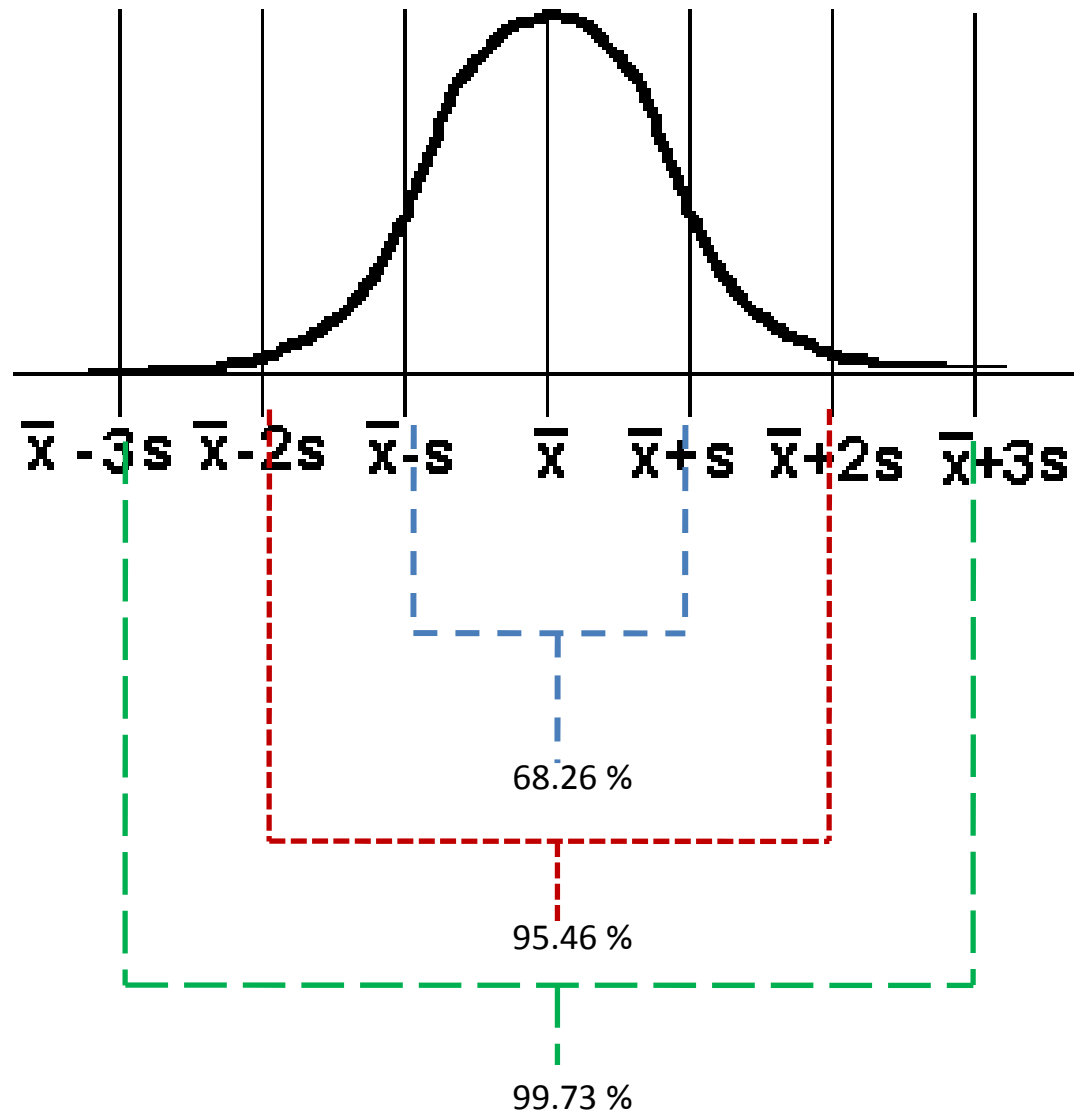


In a normal curve...

There is less than 5% chance of encountering a score greater than $\pm 2S$ from the mean

There is less than 1% chance of encountering a score greater than $\pm 3S$ from the mean

If this distribution = height, then a score outside of $3S$ is either *extremely tall* or *extremely short*, which is uncommon (improbable)



Where people get confused...

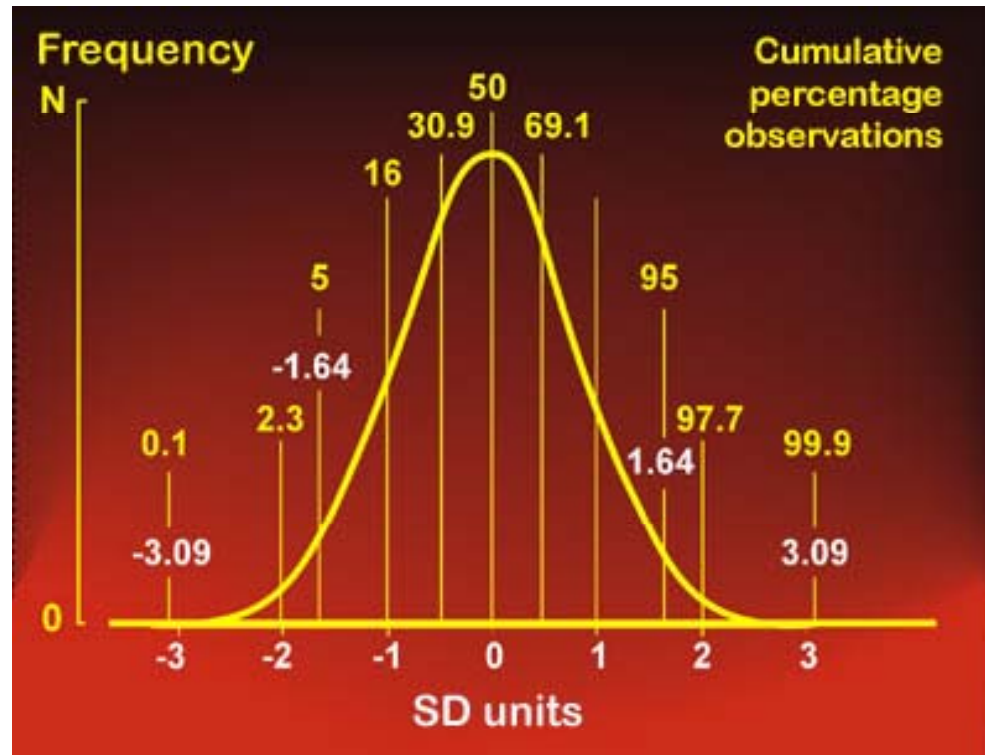
- Standard deviations not in original units
- That's why we calculate S for a variable in a sample
- So in class mean height might be 72 inches
- S might be 9 inches
- So, if $1S = 9$ inches, then
 - 68% of the people should fall between 63 and 81 inches
- **If we assume normality**

But this is, of course, relative...

- UNT men's basketball
- Mean height = 78 inches
- $S = 6$ inches
- So, 68% of the players on the team are between 72 inches and 84 inches
- If we assume normality

Or in terms of precipitation

- Mean annual precip in Denton 35" over the last 36 years
- $S = 8.5$ inches
- So, 68 percent of the years in the 36 year sample are between 27.5 and 44.5 inches in precip
- If we assume that annual precip. is normally distributed



Lowest scores

Mean

Highest scores

Standard scores (aka Z-scores)

- Let's say we want to know how many Standard deviations a particular team member is away from the mean
- We must determine the **z-score** for that player; aka "standard deviation units"
- Indicates how many standard deviations separate a particular score from the mean

$$z = \frac{(x - \bar{x})}{s}$$

- Calculated as the score value minus the mean divided by the standard deviation

Pencils

Pencil	Length (inches)	$X_i - \text{mean}$	$(X_i - \text{mean})^2$
1	10	6.3	39.69
2	4	0.3	0.09
3	2	-1.7	2.89
4	1.5	-2.2	4.84
5	1	-2.7	7.29
	Mean	Sum	$\sum(X_i - \text{mean})^2$
	3.7	0	54.8

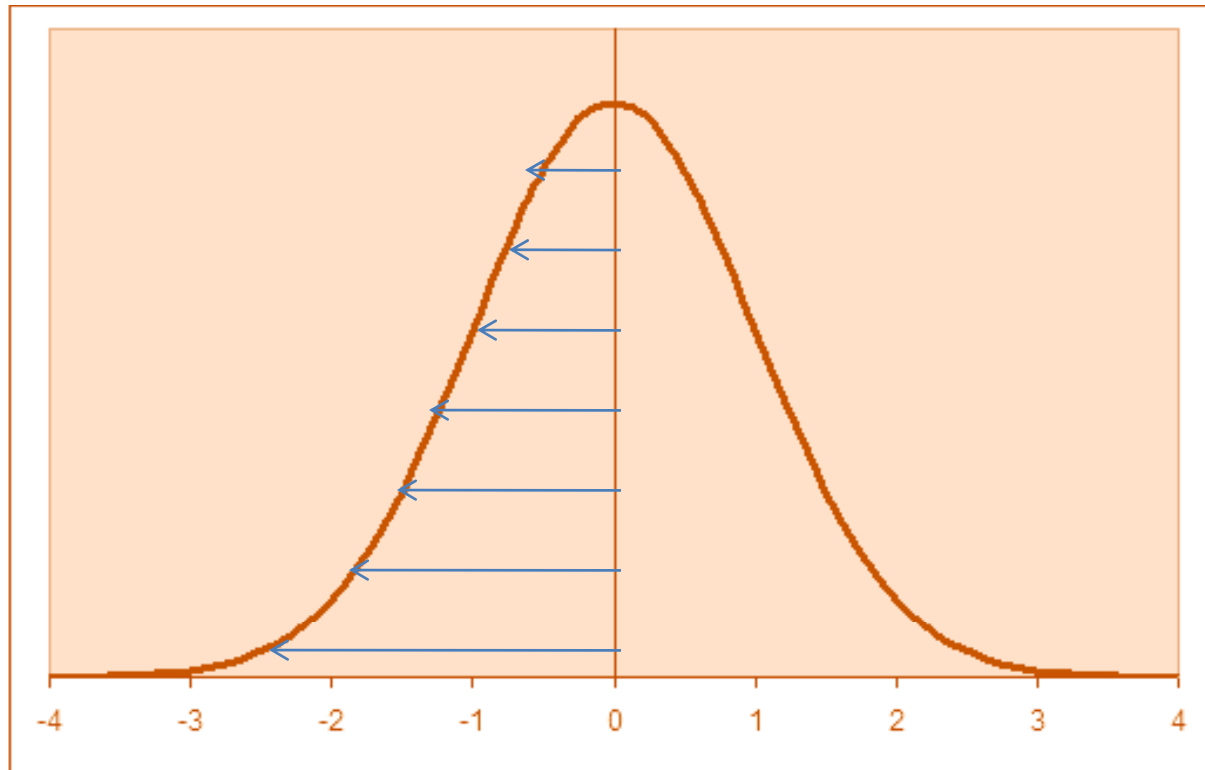
What is the variance?

What is the standard deviation?

What is the z-score for pencil 1?

Pencil 4?

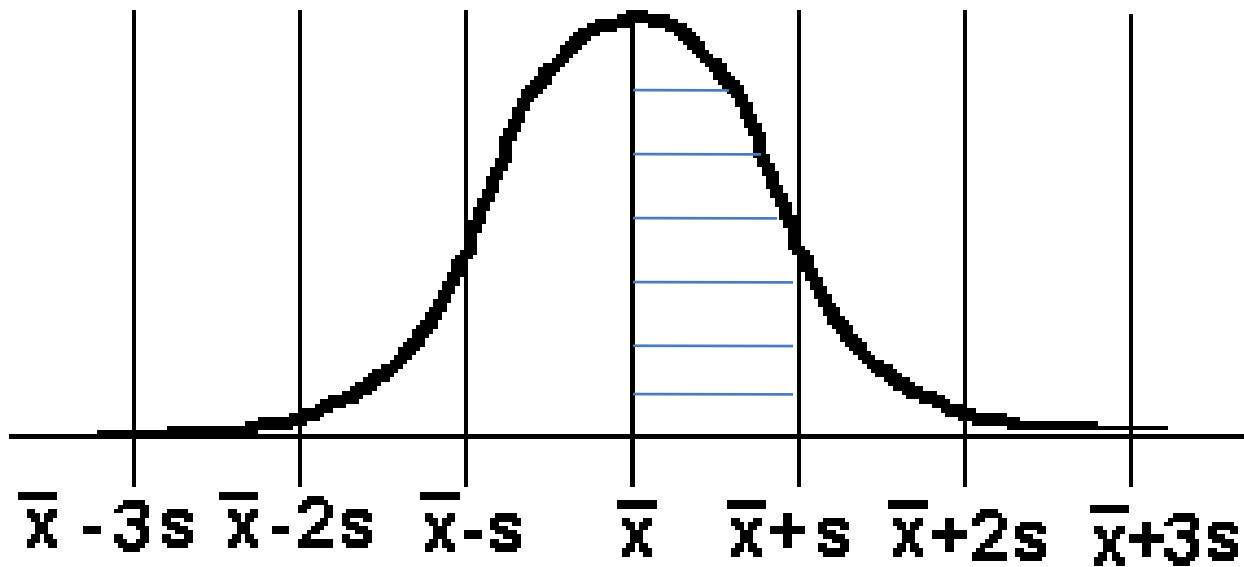
Probability of a case with score $<$ mean



Normal Curve Tables = Area

- What is the probability of encountering a case that is between the mean and +1S?
- Must find the area between the mean and 1S
- We can do this by...
 - Knowing the **z-score** for 1S ($z = 1$)
 - Using the **table**, *which is a record of area between the mean and any particular z-score*

Area from the z (normal) distribution



Summary

- 4 levels of analysis here
 - 1) raw data scores
 - 2) z-scores calculated from raw scores
 - 3) area under curve related to z-score
 - 4) area equals *probability of encounter* in a distribution

Precipitation Data

- Calculate the z-scores for each score
- Calculate Pearson's skewness
- Use z-scores to answer?
 - What is the probability of encountering a year with ≤ 32 inches in rainfall?
 - What is the probability of encountering a year with ≥ 50 inches in rainfall?
 - What is the probability of encountering a year with rainfall between 27 and 53 inches?

Why do we care?

- Most inferential tests provide a test statistic that falls in the normal distribution
- We base our conclusion on how close that test statistic is to the mean
- That is, how far is it in standard (z) scores from the mean
- **AND** how likely is it to represent the mean using probability (area)
- If it is **far out** (big z-score) the **lower the probability** it belongs with the mean.

But...

- This only works when we can assume normality...
- If samples are representative of the population, then when $n \geq 30$ we can assume normality
 - A **magic number** we will explain this week & next
- If you know that a sample is from a normally distributed population, *you can always* assume normality regardless of sample size, why?
- So, it is critical that our samples are representative...

Normality Tests in SPSS

What do these tests do?

- They compare the shape of your sample distribution to the shape of a normal curve
- The assumption is, if your sample is shaped like a normal curve, the population from which it came is normally distributed for that variable --- then you can assume normality
- A significant test means the sample distribution is not shaped like a normal curve
- Shapiro Wilks W test is the one we will use most

Shapiro Wilks W Test

- W is the test statistic
$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
- W is not significant if the variable's distribution is not different from normal
- $W \approx$ the correlation between given data and ideal normal scores
- $W = 1$ when your sample-variable data are perfectly normal (perfect H_0)
- When W is significantly smaller than 1 = non-normal (H_a is accepted)
- Shapiro-Wilk's W is recommended for small and medium samples up to $n = 2000$

Be careful

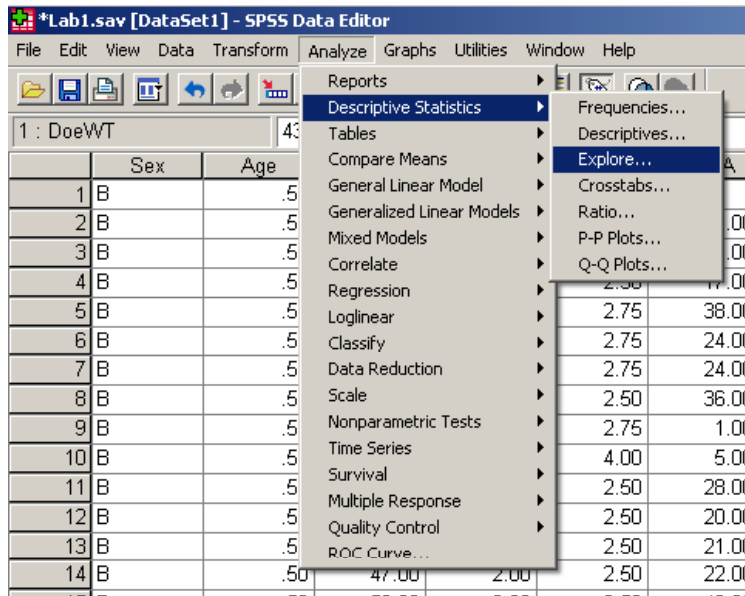
- “Do NOT use this test to say that your data are ‘normally distributed’, this assertion is quite wrong. The Shapiro-Wilk test provides evidence for certain types of ‘non-normality’ it does NOT guarantee ‘normality’”
http://www.statsdirect.com/help/parametric_methods/swt.htm
- All this test tells you is whether or not your sample looks like a normal curve
- It does not tell you that your sample is representative
- But if your sample for a variable is shaped like a normal curve, it is likely to come from a normally distributed population, if it is representative

What is a significant test result?

- When W is small enough given sample size that p is low, the results are significant
- What is p ? p is the probability of Type I error, rejecting the H_0 when it is true
- So, if p is high, we do not want to reject H_0
- If p is low, there is a low probability of Type I error

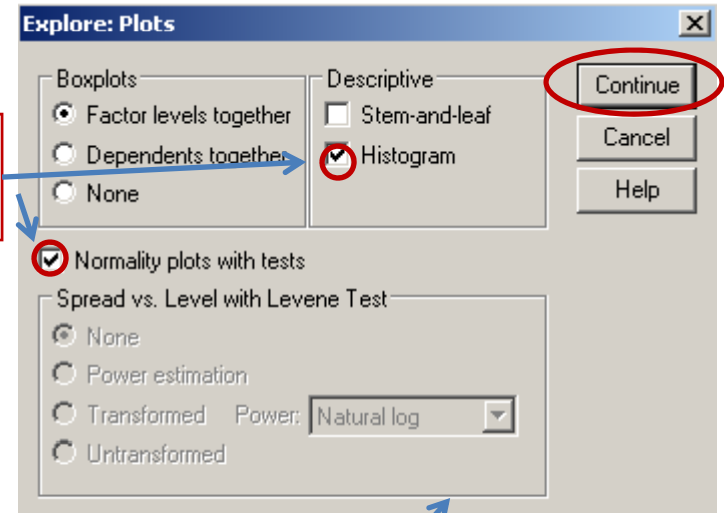
Normality test Hypotheses

- H_0 the observed distribution fits the normal distribution
- H_a the observed distribution does not fit the normal distribution
- If we accept the H_0 , we accept/assume normality

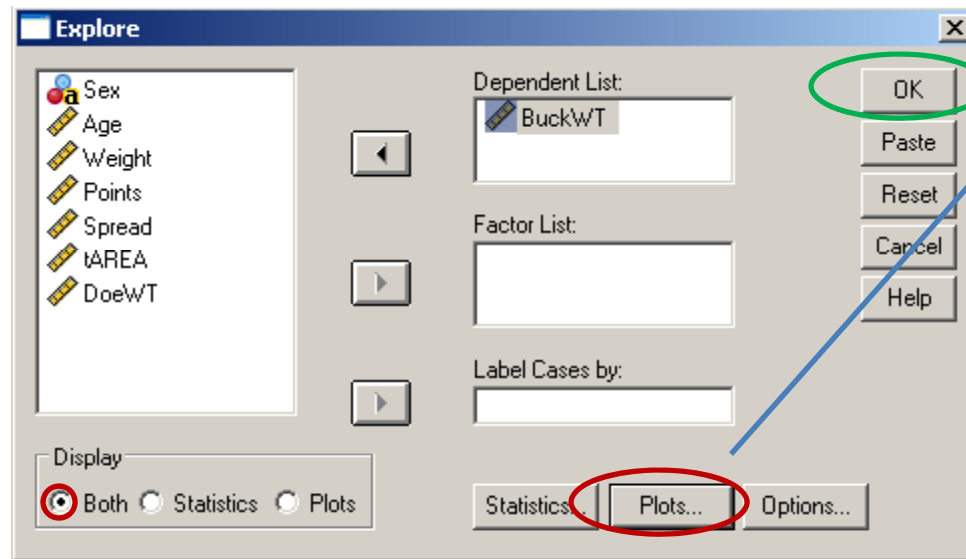


Accessing the tests

You must choose these or you will not get test results



After Explore Plots: push "OK"



Interpreting results

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BuckWT	.053	1382	.000	.986	1382	.000

a. Lilliefors Significance Correction

For tests on samples of $n = 3$ to 2000 use Shapiro Wilks; for those of $n > 2000$ use Kolmogorov-Smirnov

H_0 = normality

If you accept, then assume normality

If you reject, then do not assume normality

“Statistic” is the test statistic W for S-W, D for K-S

“Sig” is the significance for the test (aka the *p-value*)

If $p < 0.05$, reject the H_0 because the test is significant

Significance, etc.

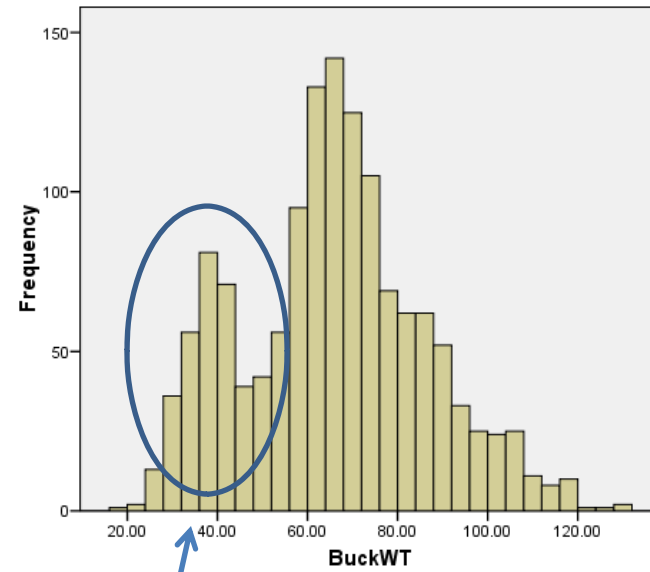
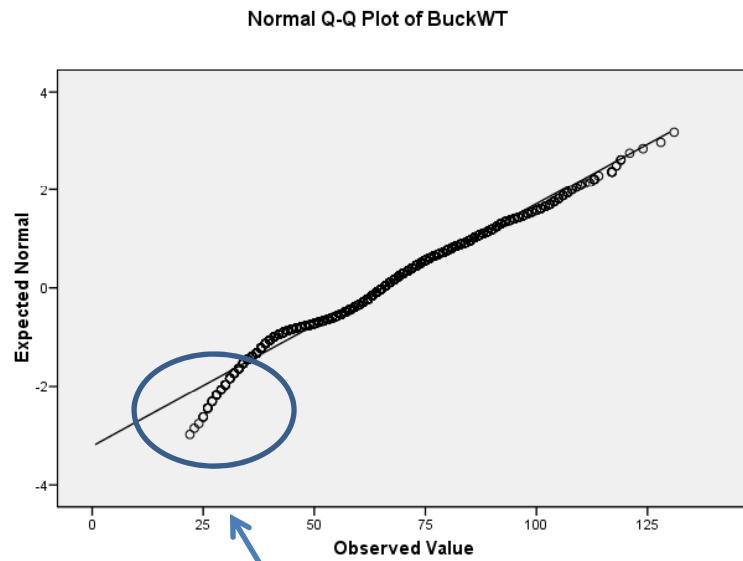
- Let's not worry about how we determine significance of a test at this point
- Simply learn the criteria and “trust it...”
- We will get back to significant results later in the class

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BuckWT	.053	1382	.000	.986	1382	.000

a. Lilliefors Significance Correction

Reject H_0



Both charts show you departure from normality at 35 to 40 pounds