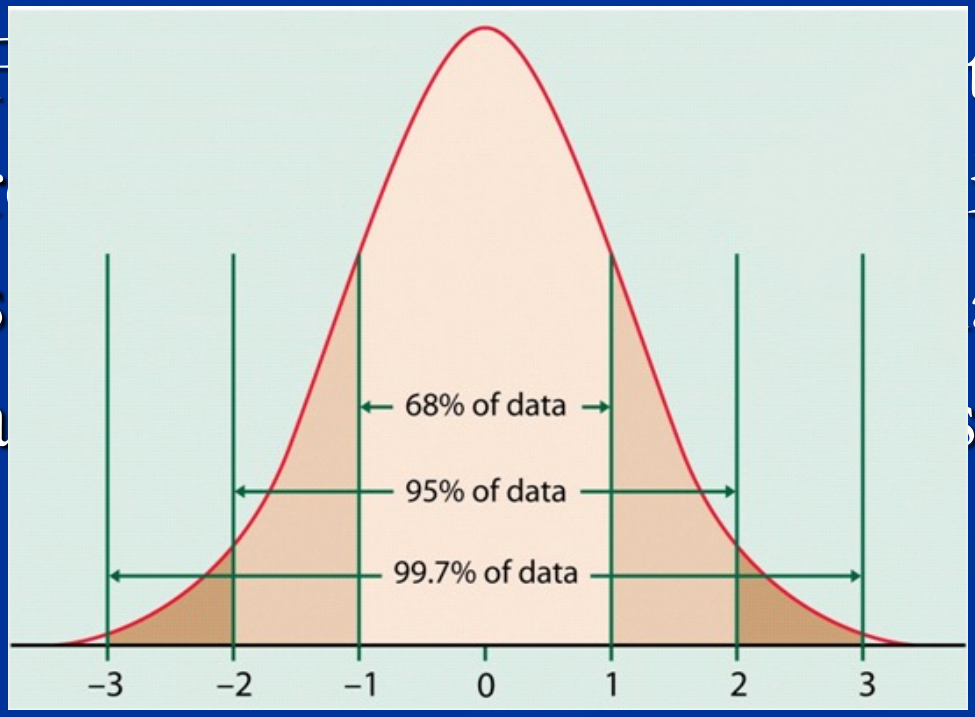
Role of Normality

- Many statistical methods require that the numeric variables we are working with have an approximate **normal distribution**.
- F ts, and r ire in some s ables are a stributed.



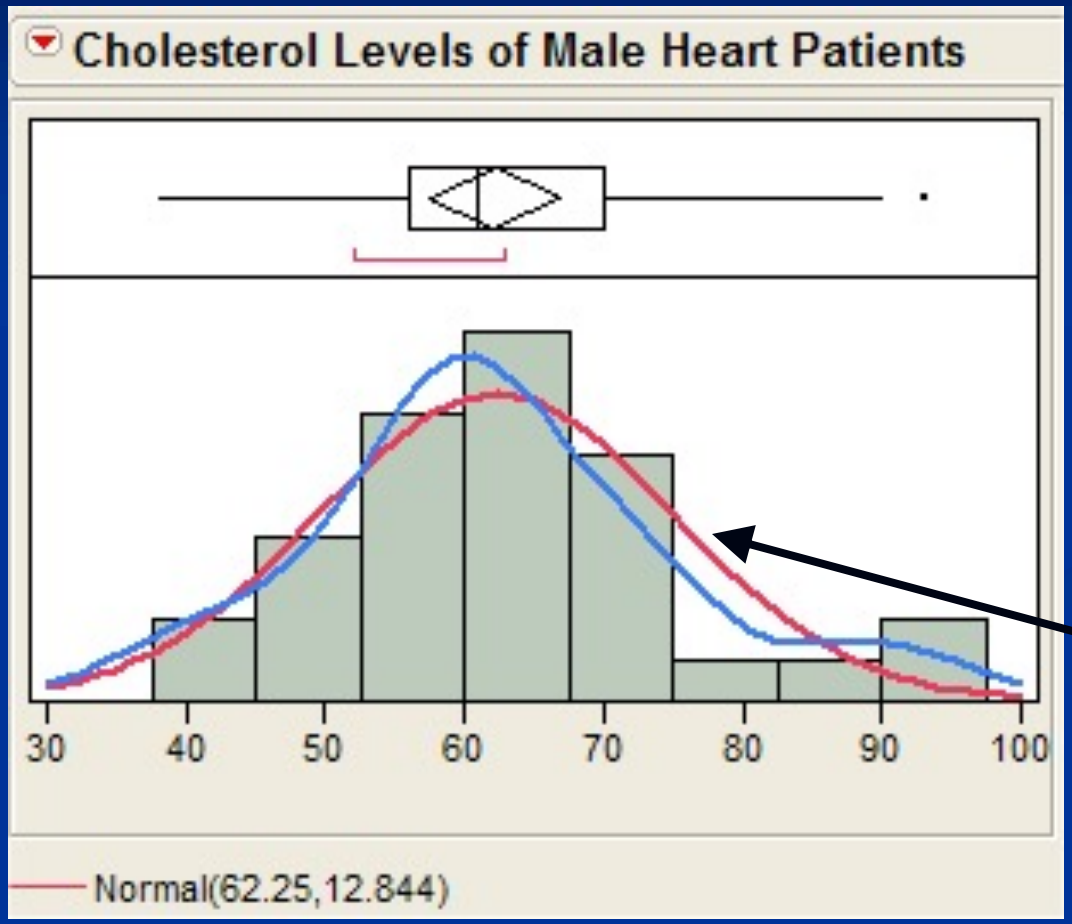
Standardized normal distribution with empirical rule percentages.

Tools for Assessing Normality

- Histogram and Boxplot
- Normal Quantile Plot
  (also called Normal Probability Plot)
- Goodness of Fit Tests

  Shapiro-Wilk Test (JMP)
  Kolmogorov-Smirnov Test (SPSS)
  Anderson-Darling Test (MINITAB)

Histograms and Boxplots

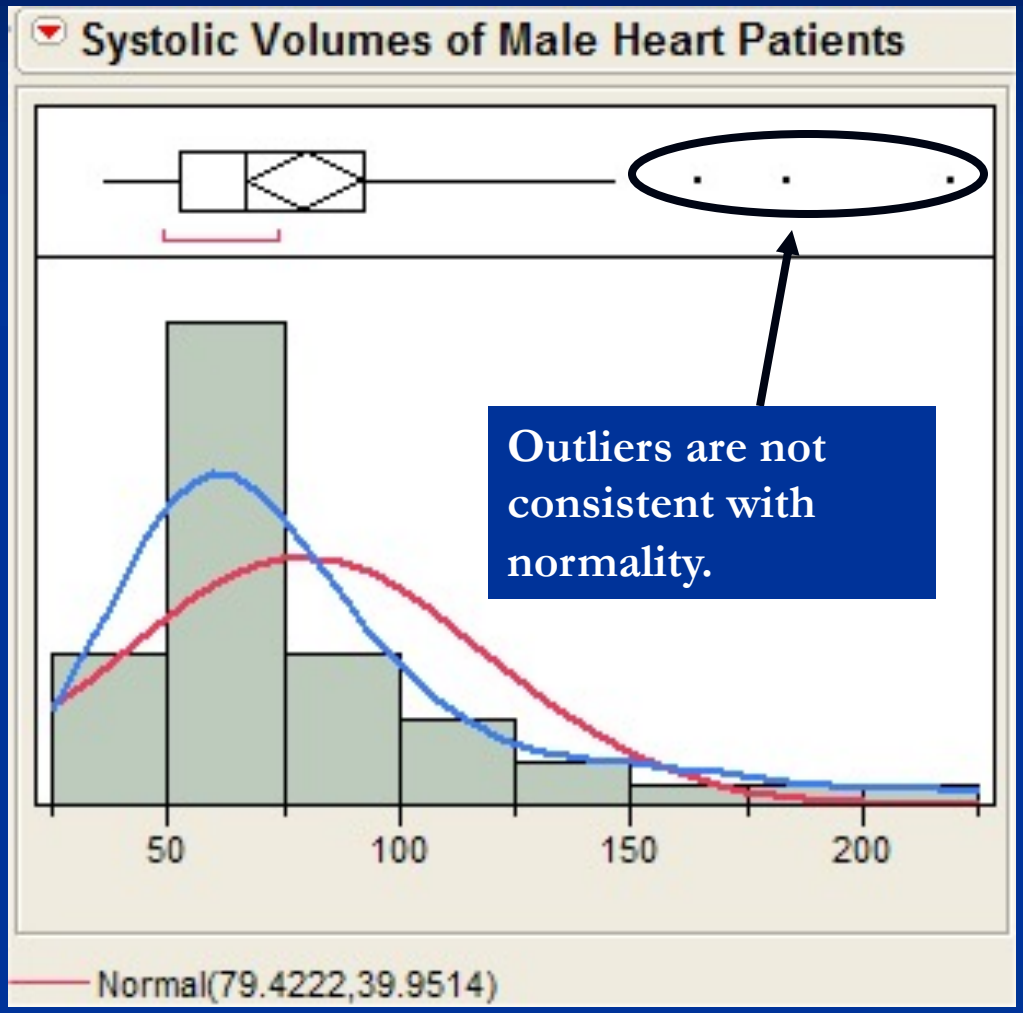


$\bar{X} = 62.25, \quad s = 12.84$

The cholesterol levels of the patients appear to be **approximately normal**, although there is some evidence of right skewness as the mean is larger than the median.

The **red** curve represents a normal distribution fit to these data and the **blue** curve the density estimate for these data, these curves should agree if our data is normally distributed.

Histograms and Boxplots



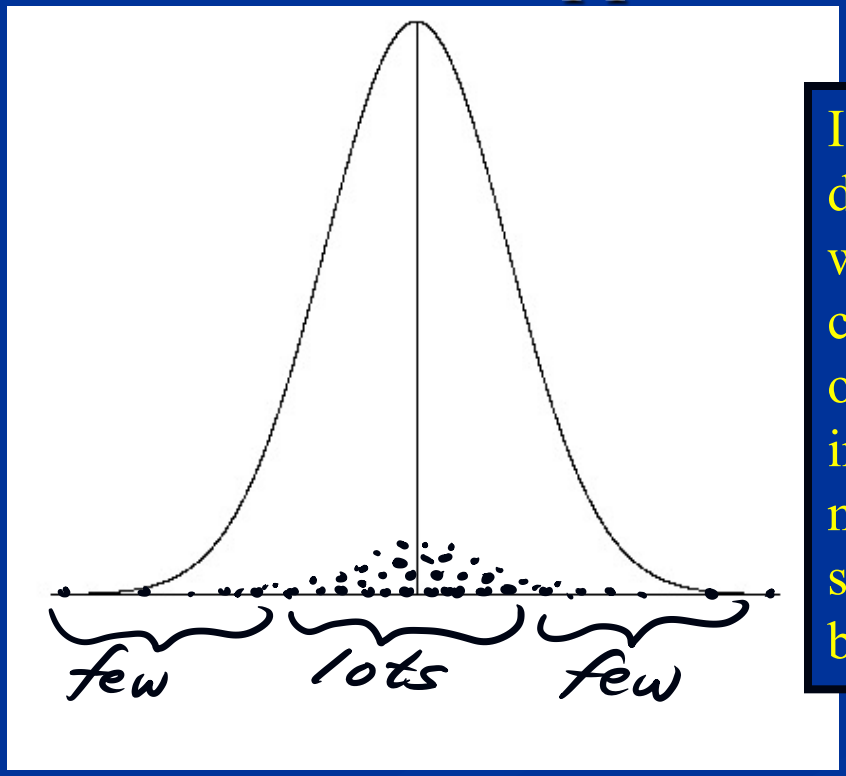
The systolic volumes of the male heart patients in this study suggest that they come from a **right skewed** population distribution.

The **red** curve represents a normal distribution fit to these data and the **blue** is the estimated density from the data which does not agree with the imposed normal.

$\bar{X} = 79.42, \quad s = 39.95$

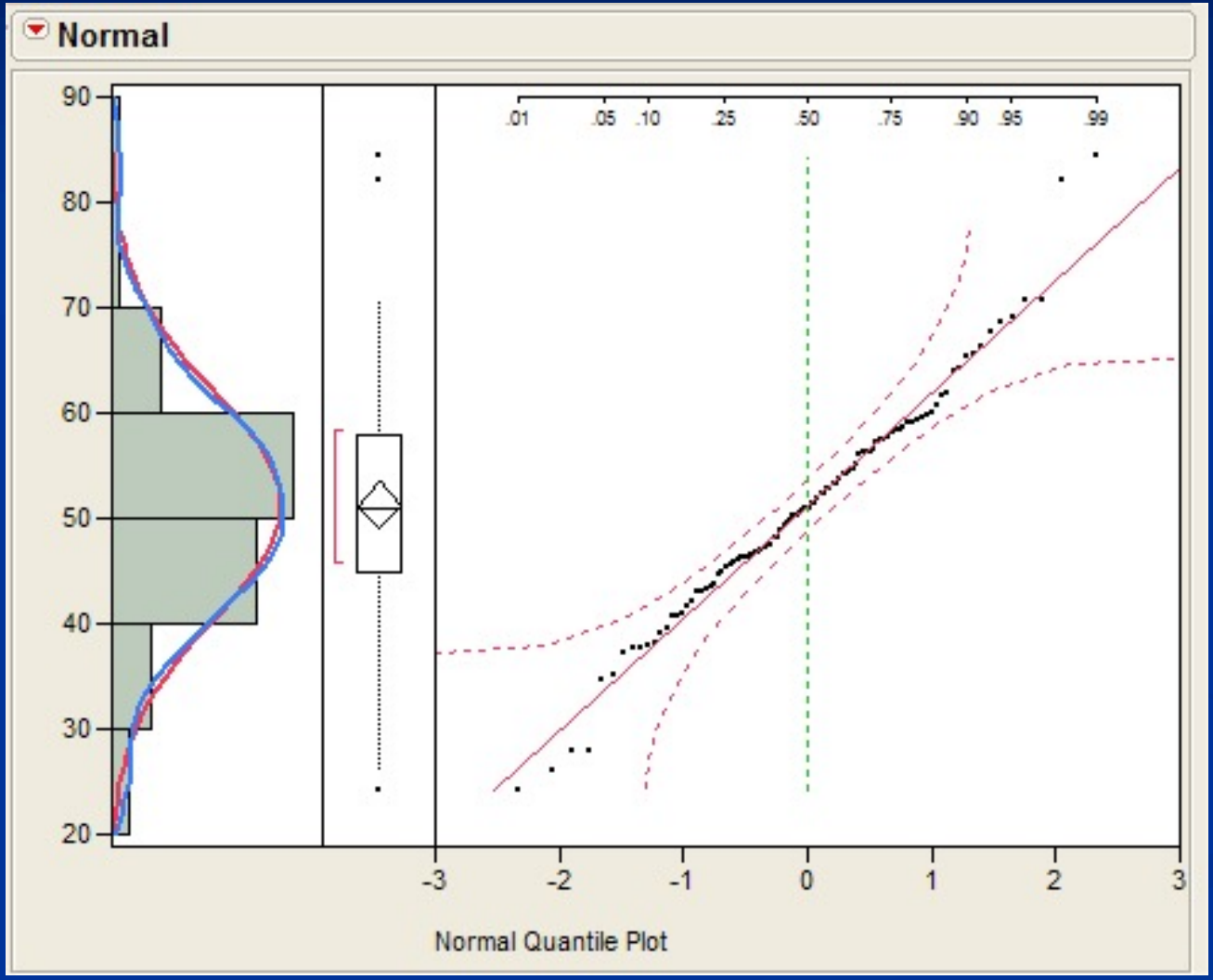
Normal Quantile Plot

- Basically compares the spacing of our data to what we would expect to see in terms of spacing if our data were approximately normal.



If our data is approximately normally distributed we should spacing similar to what I attempted to show on the normal curve on the right. Very few observations in both tails and increasingly more observations as we move towards the mean from either side. Also remember the spacing must be symmetric about the mean.

Normal Quantile Plot

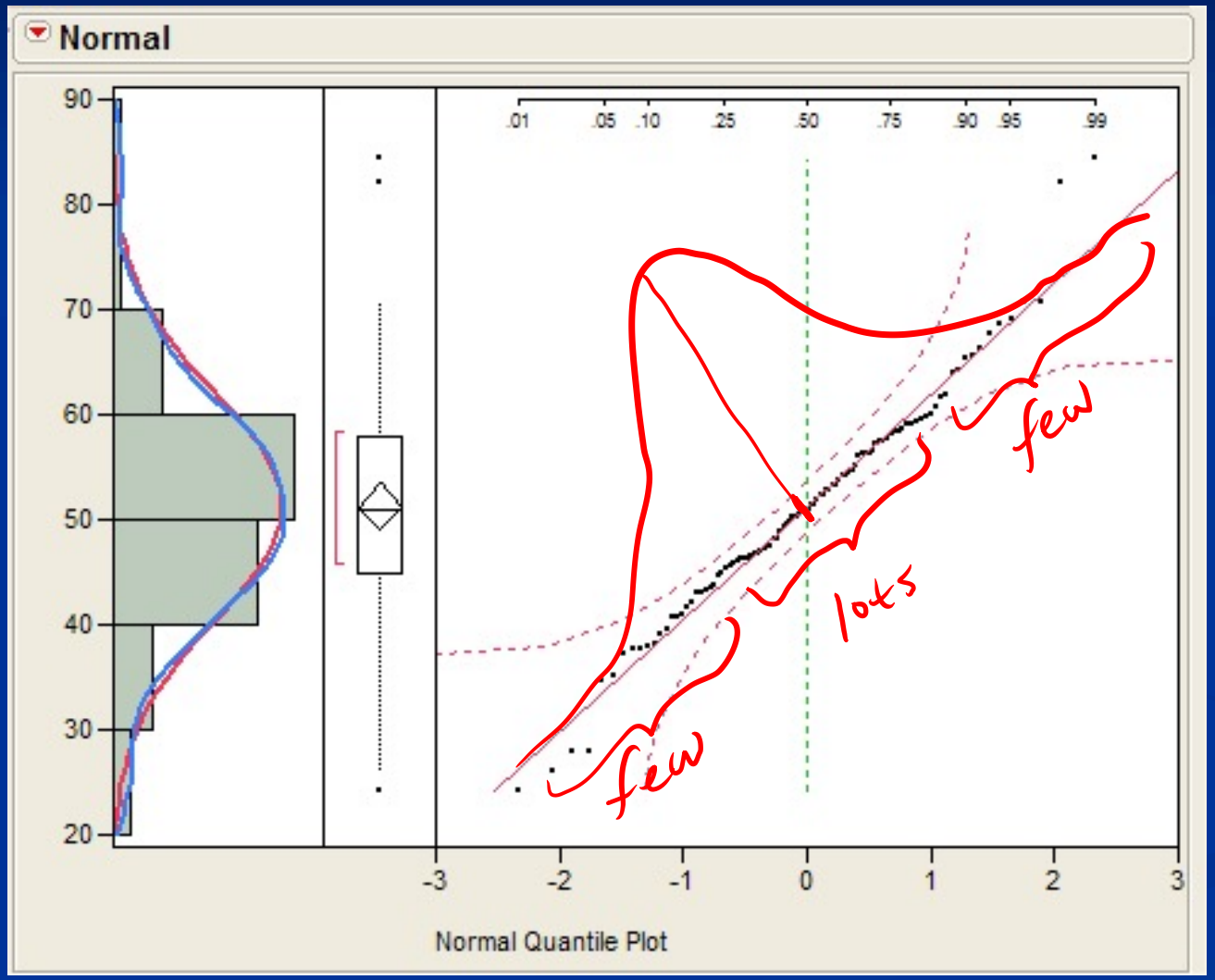


THE IDEAL PLOT:

Here is an example where the data is perfectly normal. The plot on right is a normal quantile plot with the data on the vertical axis and the expected z-scores if our data was normal on the horizontal axis.

When our data is approximately normal the spacing of the two will agree resulting in a plot with observations lying on the reference line in the normal quantile plot. The points should lie within the dashed lines.

Normal Quantile Plot

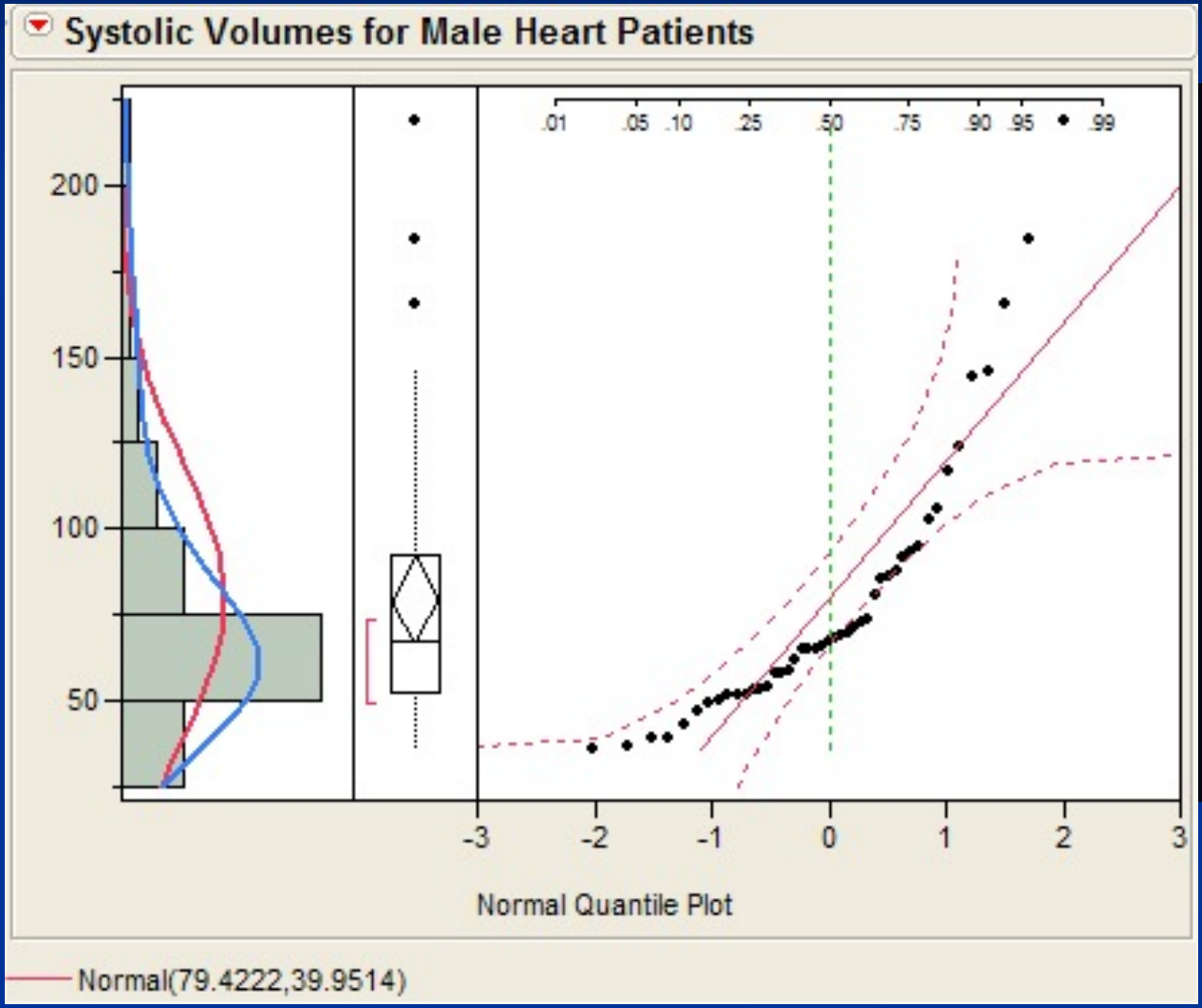


THE IDEAL PLOT:

Here is an example where the data is perfectly normal. The plot on right is a normal quantile plot with the data on the vertical axis and the expected z-scores if our data was normal on the horizontal axis.

When our data is approximately normal the spacing of the two will agree resulting in a plot with observations lying on the reference line in the normal quantile plot. The points should lie within the dashed lines.

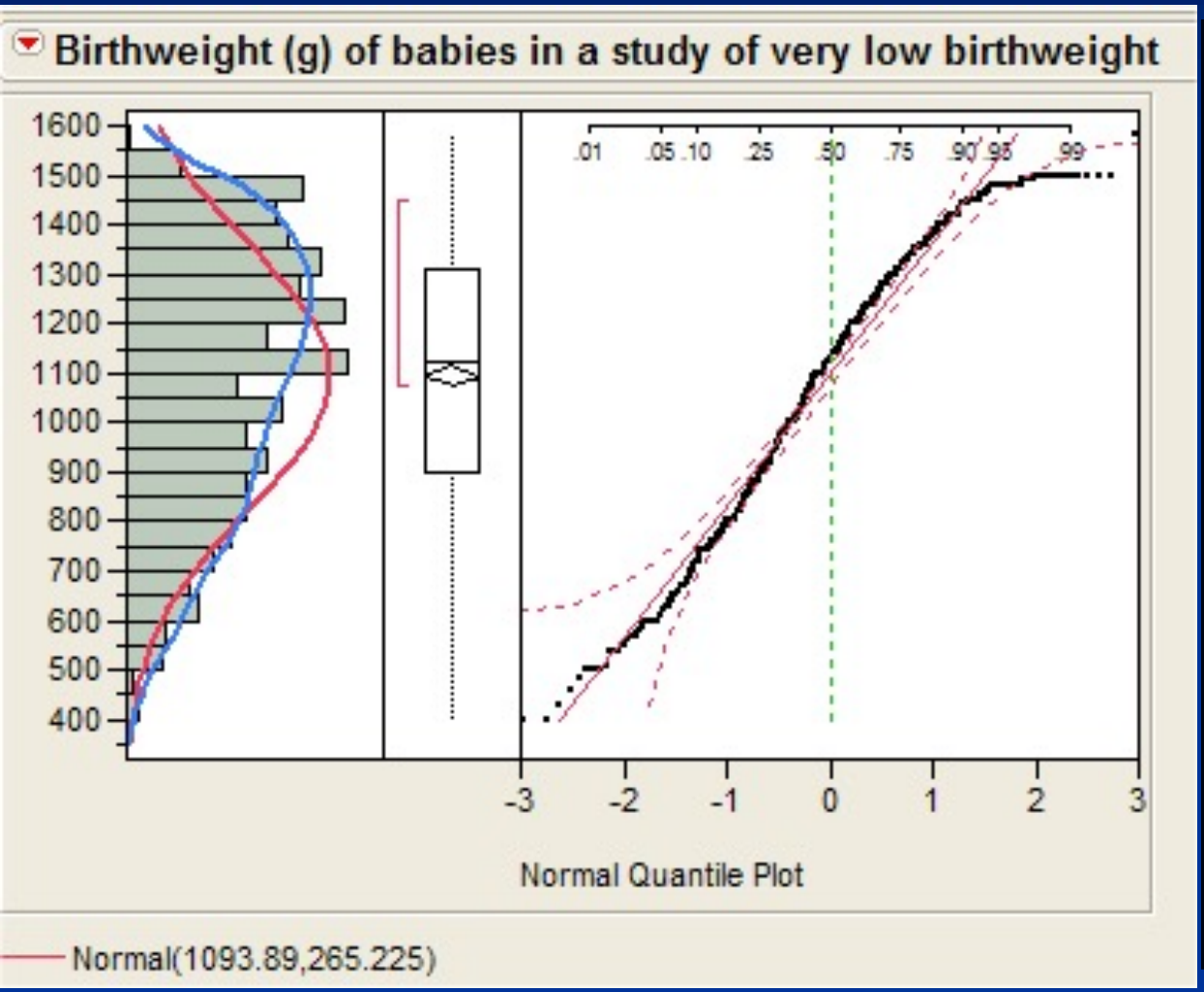
Normal Quantile Plot (right skewness)



The systolic volumes of the male heart patients are clearly **right skewed**.

When the data is plotted vs. the expected z-scores the normal quantile plot shows right skewness by a **upward bending** curve

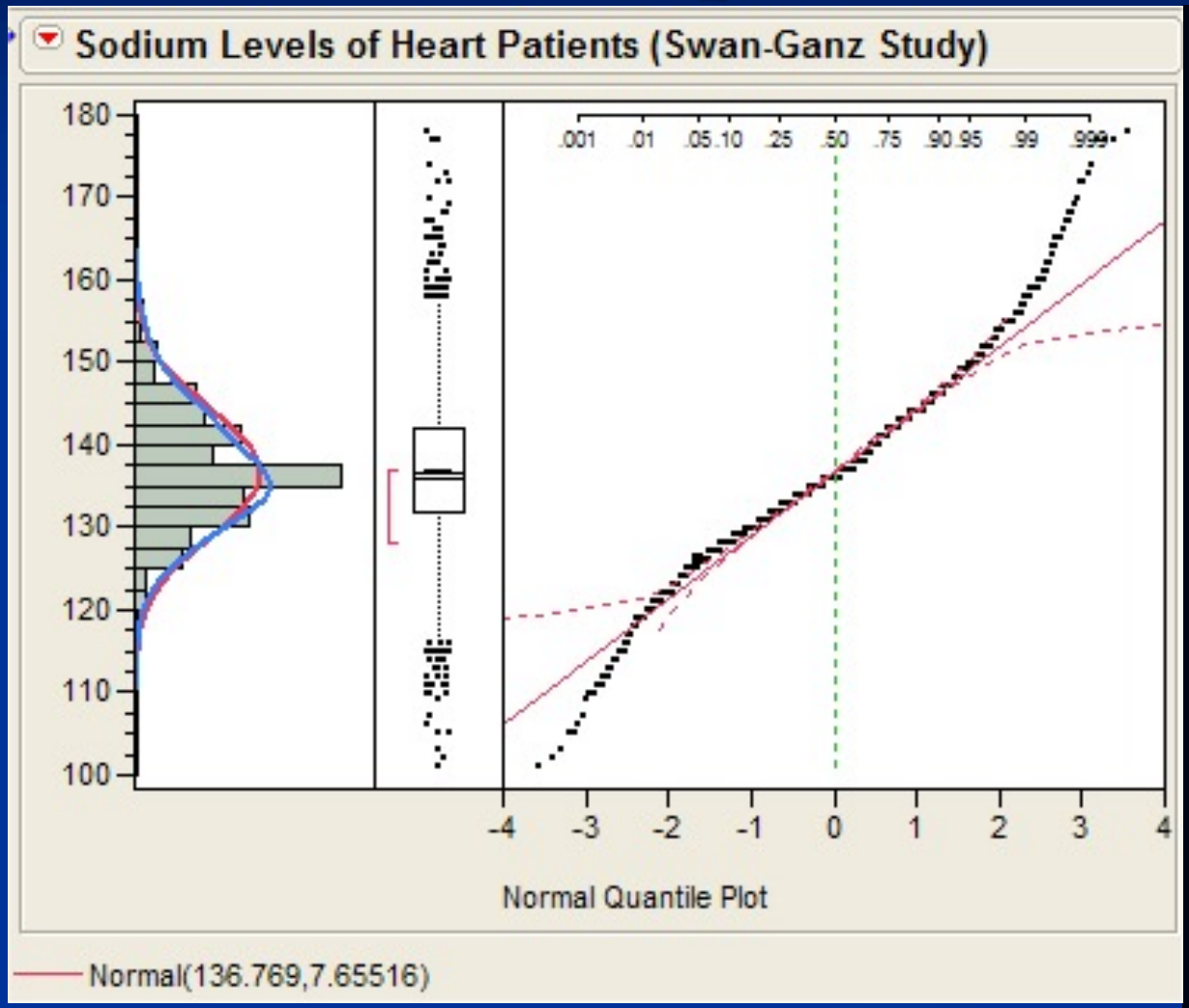
Normal Quantile Plot (left skewness)



The distribution of birthweights from this study of very low birthweight infants is **skewed left**.

When the data is plotted vs. the expected z-scores the normal quantile plot shows left skewness by a **downward bending** curve.

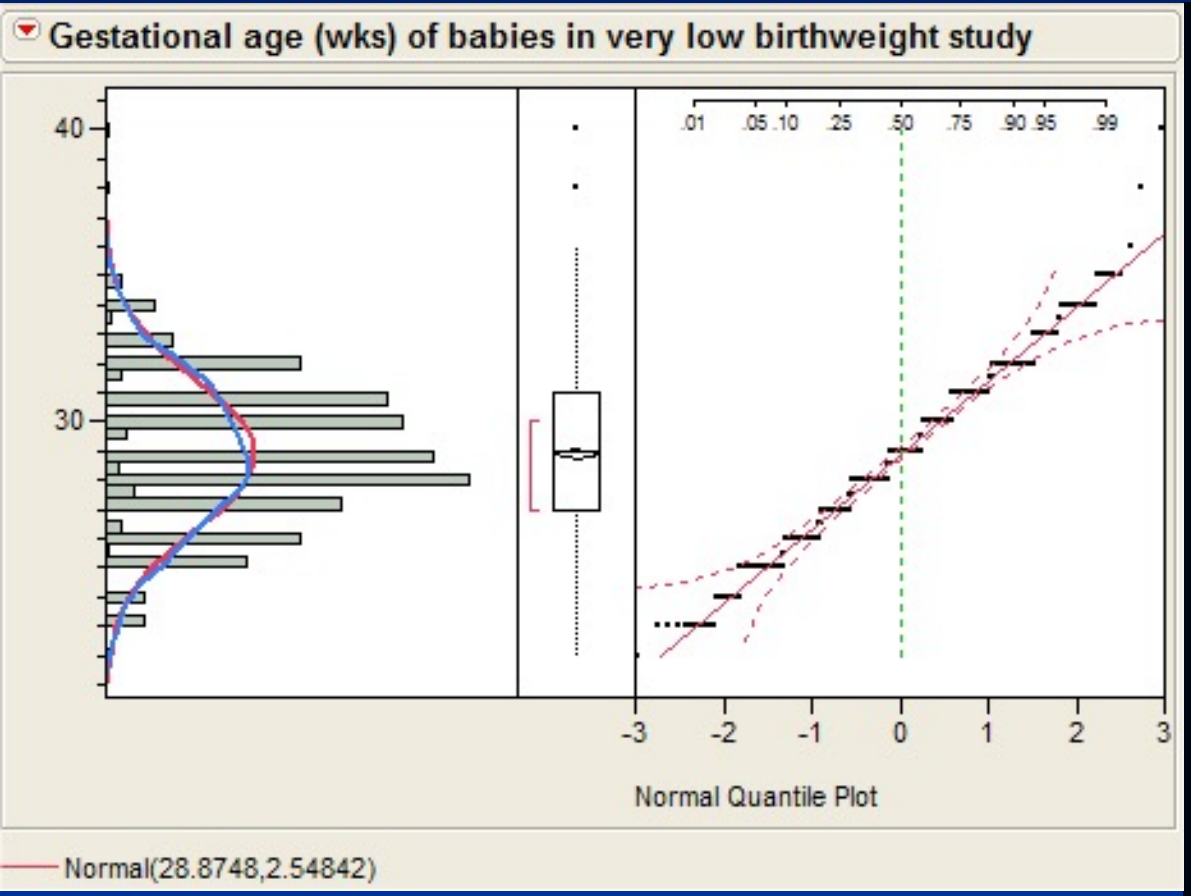
Normal Quantile Plot (leptokurtosis)



The distribution of sodium levels of patients in this right heart catheterization study has **heavier tails** than a normal distribution (i.e, leptokurtosis).

When the data is plotted vs. the expected z-scores the normal quantile plot there is an **"S-shape"** which indicates **kurtosis**.

Normal Quantile Plot (discrete data)



Although the distribution of the gestational age data of infants in the very low birthweight study is approx. normal there is a **"staircase"** appearance in normal quantile plot.

This is due to the **discrete coding** of the gestational age which was recorded to the nearest week or half week.

Normal Quantile Plots

IMPORTANT NOTE:

- If you plot **DATA vs. NORMAL** as on the previous slides then:

  downward bend = left skew

  upward bend = right skew

- If you plot **NORMAL vs. DATA** then:

  downward bend = right skew

  upward bend = left skew

Tests of Normality

There are several different tests that can be used to test the following hypotheses:

H_o: The distribution is normal

H_A: The distribution is NOT normal

Common tests of normality include:

Shapiro-Wilk **Kolmogorov-Smirnov**

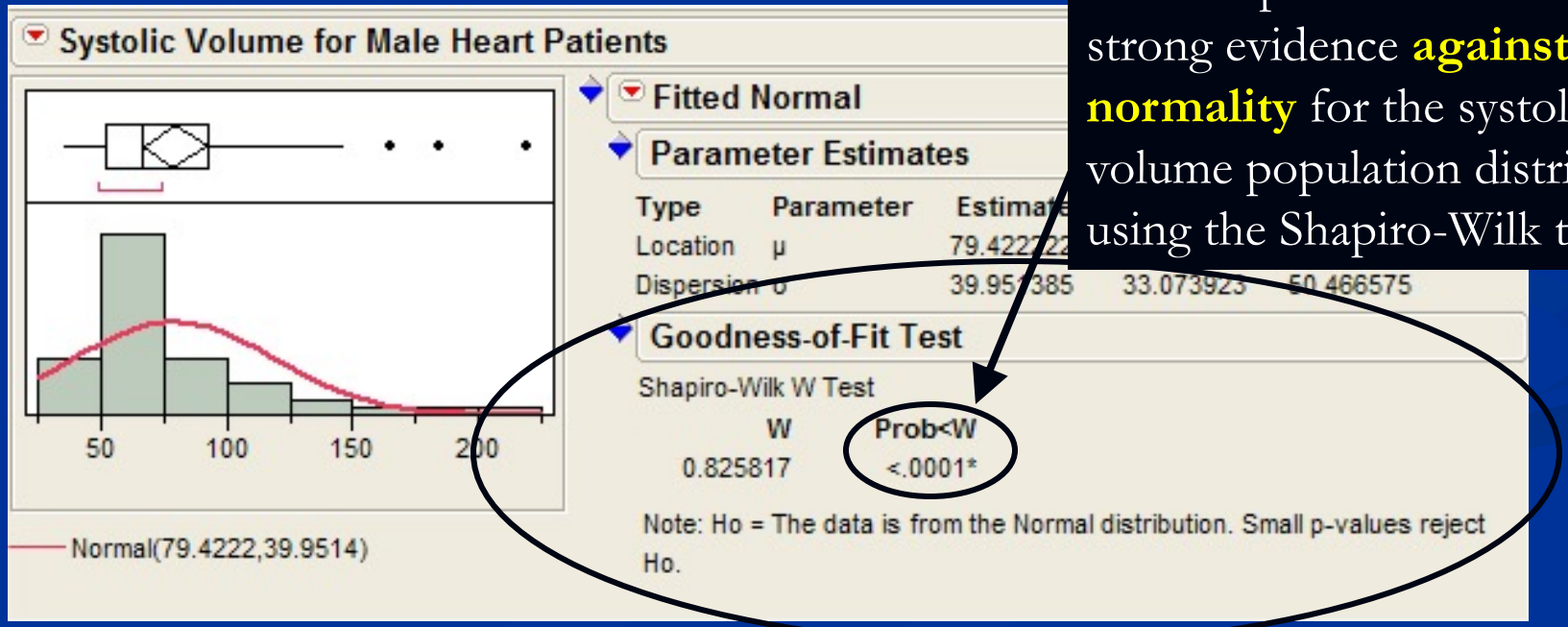
Anderson-Darling **Lillefor's**

Problem: THEY DON'T ALWAYS AGREE!!

Tests of Normality

H_o: The distribution of systolic volume is normal

H_A: The distribution of systolic volume is NOT normal



Because p < .0001 we have strong evidence **against normality** for the systolic volume population distribution using the Shapiro-Wilk test.

Tests of Normality

H_o: The distribution of systolic volume is normal

H_A: The distribution of systolic volume is NOT normal

One-Sample Kolmogorov-Smirnov Test

		SYSVOL
N		45
Normal Parameters[a,b]	Mean	79.4222
	Std. Deviation	[illegible]
Most Extreme Differences	Absolute	.198
	Positive	.198
	Negative	-.139
Kolmogorov-Smirnov Z		1.331
Asymp. Sig. (2-tailed)		.058

a. Test distribution is Normal.

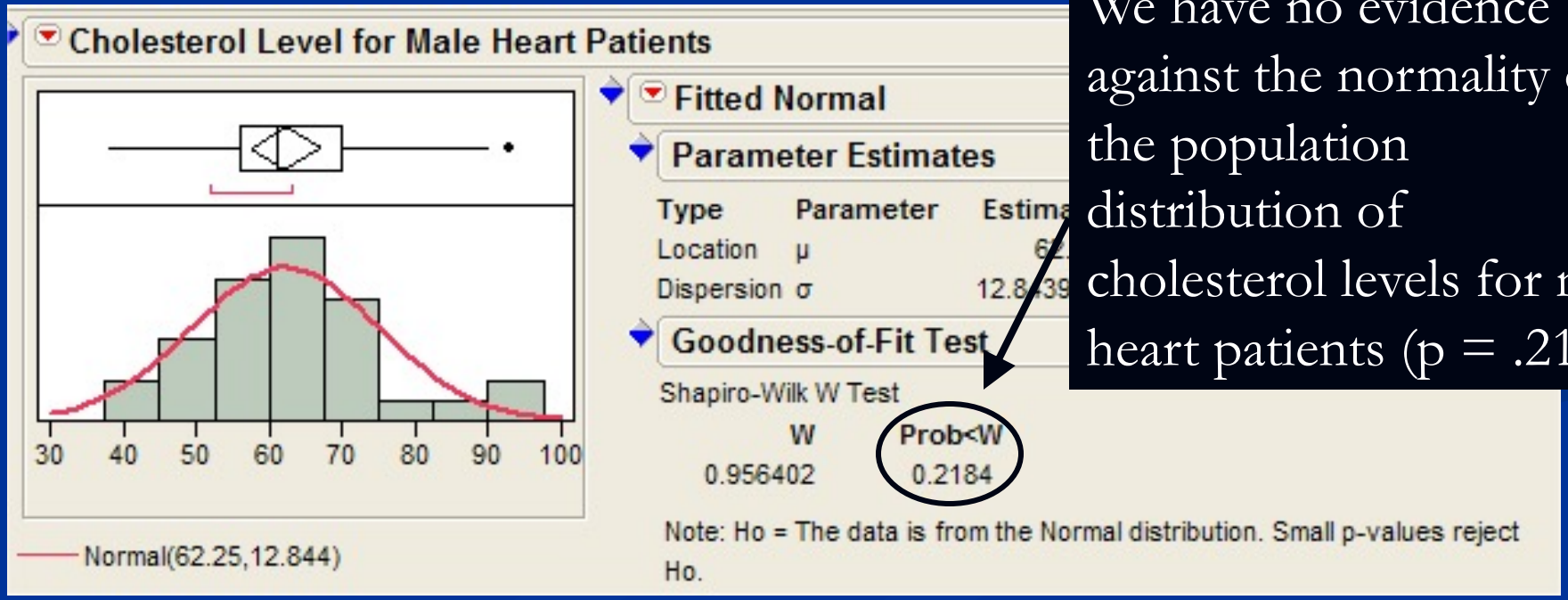
b. Calculated from data.

We do not have evidence at the $\alpha = .05$ level against the normality of the population systolic volume distribution when using the Kolmogorov-Smirnov test from SPSS.

Tests of Normality

H_o: The distribution of cholesterol level is normal

H_A: The distribution of cholesterol level is NOT normal



We have no evidence against the normality of the population distribution of cholesterol levels for male heart patients (p = .2184).

Transformations to Improve Normality (removing skewness)

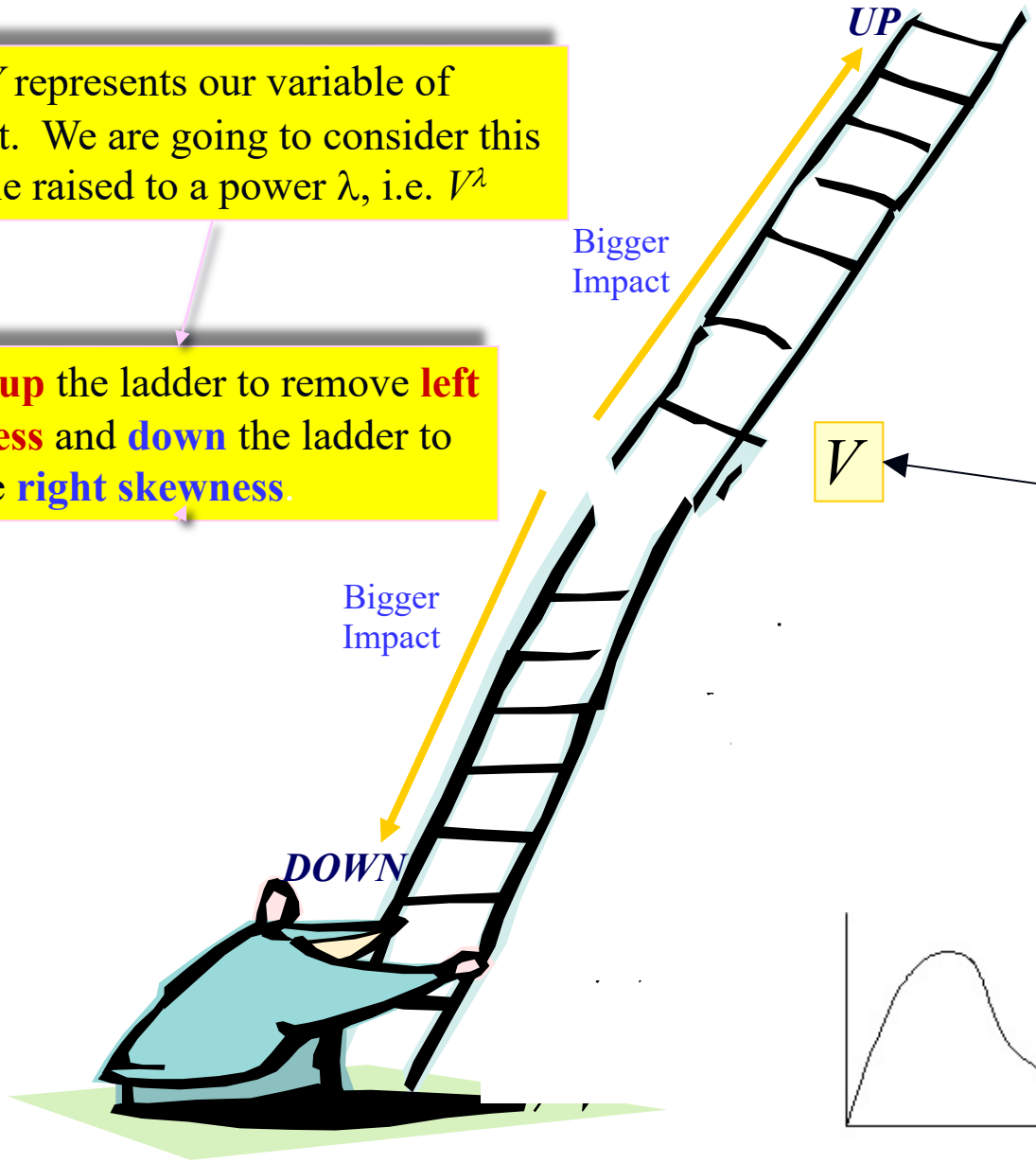
Many statistical methods require that the numeric variables you are working with have an approximately normal distribution.

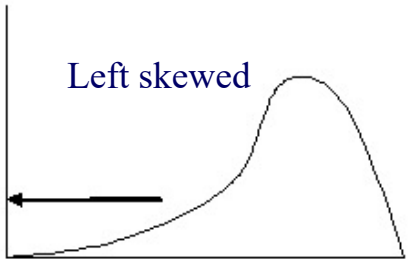
Reality is that this is often times not the case. One of the most common departures from normality is skewness, in particular, **right skewness**.

Tukey's Ladder of Powers

Here V represents our variable of interest. We are going to consider this variable raised to a power λ, i.e. V^{λ}

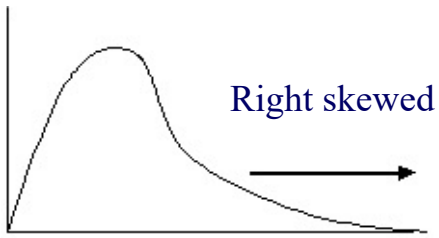
We go **up** the ladder to remove **left skewness** and **down** the ladder to remove **right skewness**.





V

Middle rung:
No transformation
($\lambda = 1$)



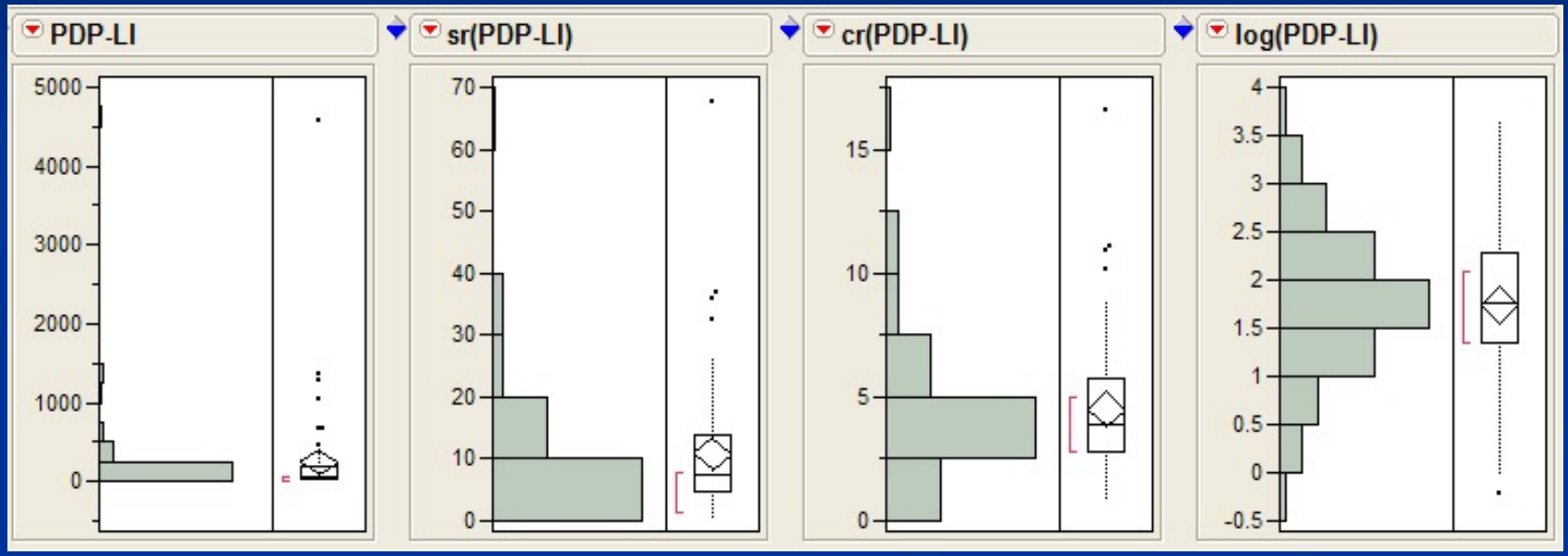
Tukey's Ladder of Powers

- To remove **right skewness** we typically take the square root, cube root, logarithm, or reciprocal of a the variable etc., i.e. $V^{.5}$, $V^{.333}$, $log_{10}(V)$ (think of V^0) , V^{-1}, *etc.*

- To remove **left skewness** we raise the variable to a power greater than 1, such as squaring or cubing the values, i.e. V^2, V^3, etc.

Removing Right Skewness

Example 1: PDP-LI levels for cancer patients

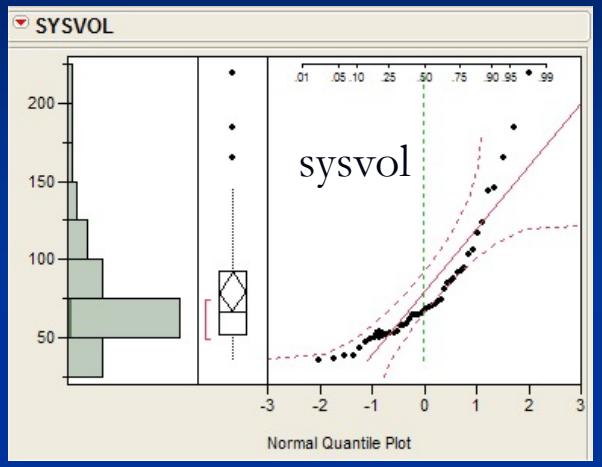


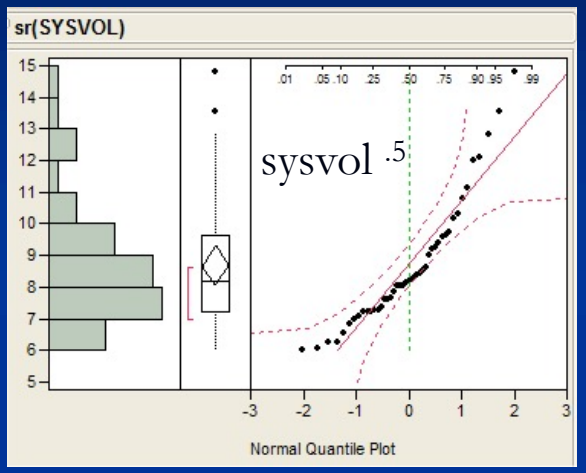
$PDP - LI$ $\sqrt{PDP-LI}$ $\sqrt[3]{PDP-LI}$ $\log_{10}(PDP-LI)$

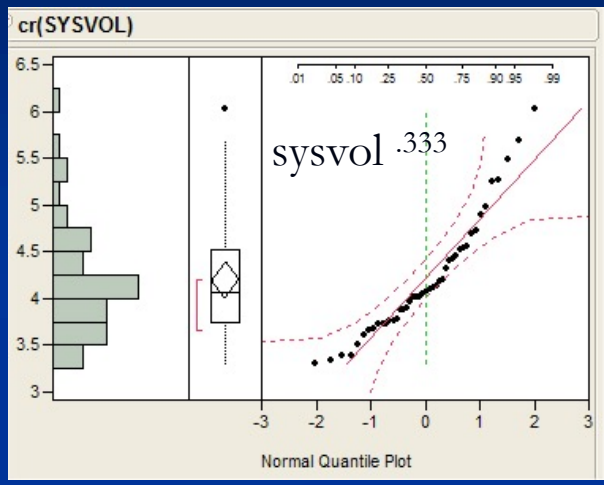
In the log base 10 scale the PDP-LI values are approximately normally distributed.

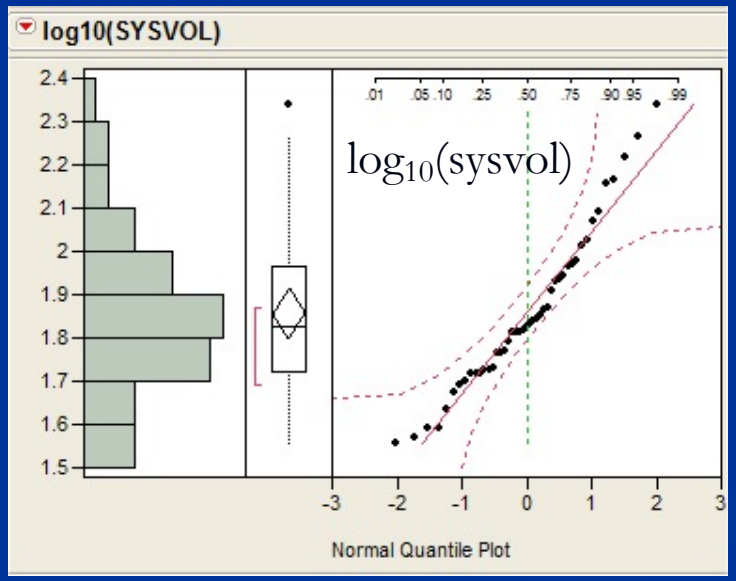
Removing Right Skewness

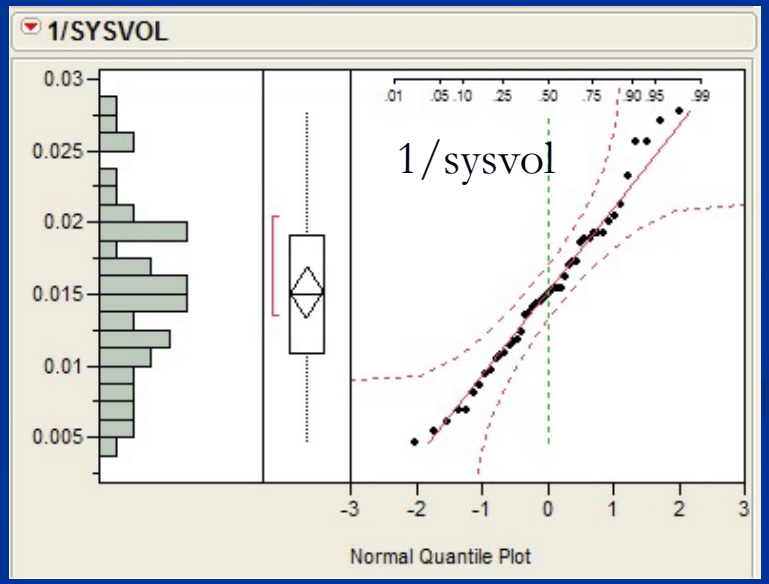
Example 2: Systolic Volume for Male Heart Patients





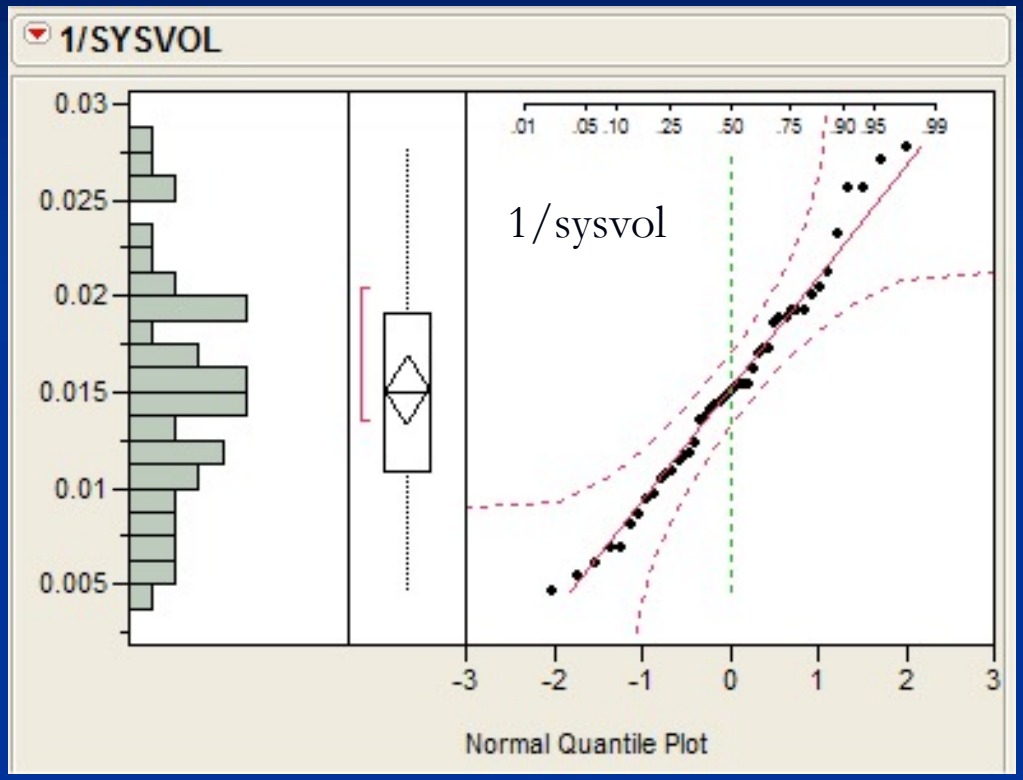






Removing Right Skewness

Example 2: Systolic Volume for Male Heart Patients



Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	0.0150757	0.0133388	0.0168126
Dispersion	σ	0.0057813	0.0047861	0.007303

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.977818	0.5340

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

The reciprocal of systolic volume is approximately normally distributed and the Shapiro-Wilk test provides no evidence against normality (p = .5340).

CAUTION: The use of the reciprocal transformation reorders the data in the sense that the largest value becomes the smallest and the smallest becomes the largest after transformation. The units after transformation may or may not make sense, e.g. if the original units are mg/ml then after transformation they would be ml/mg.

Try qqplot.m in Matlab!

And randn.m!

Nonparametric Methods (distribution-free methods)

- Same general procedure:
  1. Make some claim about the underlying populations in the form of a null hypothesis.
  2. Calculate the value of the test statistic using data contained in a random sample of observations.
  3. Depending on the magnitude of the statistic, either reject or do not reject the null hypothesis.

The Sign Test

- Compare two samples of observations when the populations are not independent.
- (like paired t-test, focuses on difference in values for each pair.)
- Population of differences need not be normally distributed.
- Null hypothesis: in the underlying population of differences among pairs,

  median difference = 0.

The Sign Test

- Example: REE: resting energy expenditure (REE).
- Compare patients with cystic fibrosis and healthy individuals matched on age, sex, height, and weight.
- Null hypothesis: median difference = 0.
- For 2-sided test, the alternative hypothesis: median difference ≠ 0

The Sign Test

	REE (kcal/day)			
Pair	CF	Healthy	Difference	Sign
1	1153	996	157	+
2	1132	1080	52	+
3	1165	1182	-17	-
4	1460	1452	8	+
5	1634	1162	472	+
6	1493	1619	-126	-
7	1358	1140	218	+
8	1453	1123	330	+
9	1185	1113	72	+
10	1824	1463	361	+
11	1793	1632	161	+
12	1930	1614	316	+
13	2075	1836	239	+

The Sign Test

- Differences of exactly 0 are excluded.
- D: number of plus signs
- Plus sign “success”
- Bernoulli random variable with probability of success p = 0.5
- D is a binomial random variable with parameters n and p.

A **Bernoulli random variable** is the simplest kind of random variable. It can take on two values, 1 and 0. It takes on a 1 if an experiment with probability *p* resulted in success and a 0 otherwise. Some example uses include a coin flip, a random binary digit, whether a disk drive crashed, and whether someone likes a Netflix movie.

A **binomial random variable** is random variable that represents the number of successes in *n* successive independent trials of a Bernoulli experiment. Some example uses include the number of heads in *n* coin flips, the number of disk drives that crashed in a cluster of 1000 computers, and the number of advertisements that are clicked when 40,000 are served.

The Sign Test

- Mean number of plus signs in a sample of size n is: np = n/2
- Standard deviation = sqrt(np(1-p)) = sqrt(n/4)
- If D is either much larger or much smaller than n/2, we will reject H_0.
- Test statistic: $z_+ = (D – (n/2))/\text{sqrt}(n/4)$
- If the null hypothesis is true and the sample size n is large, then z_+ follows an approximate normal distribution with mean 0 and std. dev. = 1.

The Sign Test

- For our example, D=11 plus signs.
- Also, n/2 = 13/2 = 6.5
- sqrt(n/4) = 1.80
- Thus, $z_+ = (11 - 6.5)/1.80 = 2.50$
- Area under standard normal curve to the right of z = 2.50 and to the left of z = -2.50 is:

$$p = 2(0.006) = 0.012$$

Since $p<0.05$, we reject the null hypothesis, conclude that median difference among pairs is $\neq 0$.

Wilcoxon Signed-Rank Test

- The sign test ignores the magnitude of the difference.
- Wilcoxon Signed-Rank Test can also be used to compare 2 samples from populations that are not independent.
- Focuses on difference in values for each pair of observations (like the sign test).
- Population of difference does not need to be normally distributed.

Wilcoxon Signed-Rank Test

- Null hypothesis: in the underlying population of differences among pairs, the median difference = 0.
- Example: FVC: forced vital capacity.
- Volume of air that person can expel from lungs in 6 seconds.
- Test drug amiloride as therapy for cystic fibrosis, over 2.5 weeks.

Wilcoxon Signed-Rank Test

1. Select random sample of n pairs of observations.
2. Calculate the difference for each pair of observations. Ignoring sign of difference, rank absolute values from smallest to largest. Difference = 0 is eliminated from analysis. Tied observations are given an average rank (e.g., two pairs = 1.5).
3. Assign each rank a + or – sign, depending on the sign of the difference.

Wilcoxon Signed-Rank Test

Subject	Placebo	Drug	Difference	Rank	Signed rank		
1	224	213	11	1	1		
2	80	95	-15	2		-2	
3	75	33	42	3	3		
4	541	440	101	4	4		
5	74	-32	106	5	5		
6	85	-28	113	6	6		
7	293	445	-152	7		-7	
8	-23	-178	155	8	8		
9	525	367	158	9	9		
10	-38	140	-178	10		-10	
11	508	323	185	11	11		
12	255	10	245	12	12		
13	525	65	460	13	13		
14	1023	343	680	14	14		
					86	-19	totals

Wilcoxon Signed-Rank Test

4. Compute the sum of the positive ranks and the sum of the negative ranks. Ignoring the signs, denote the smaller sum by T.

- Under the null hypothesis (median diff = 0) we expect sample to have approximately equal number of positive and negative ranks.
- Sum of the positive ranks should be comparable in magnitude to sum of negative ranks.

Wilcoxon Signed-Rank Test

- Test statistic: $z_T = (T - \mu_T)/\sigma_T$
- Where $\mu_T = n(n+1)/4$ is the mean sum of ranks.
- $\sigma_T = \text{sqrt}(n(n+1)(2n+1)/24)$ is the std. dev.
- If H_0 is true and the sample size n is large enough, $z_T = (T - \mu_T)/\sigma_T$

  follows an approximately normal distribution

  with mean 0 and std. dev. 1.

Wilcoxon Signed-Rank Test

- For our example, T=19.
- $\mu_T = 52.5,\ \sigma_T = 15.93$
- Therefore, $z_T = -2.10$
- Area under standard normal curve to the left of z = -2.10 and to the right of z = 2.10 is:

  $p = 2(0.018) = 0.036$
- Since $p < 0.05$, we reject the null hypothesis and conclude that the median difference is $\neq 0$.

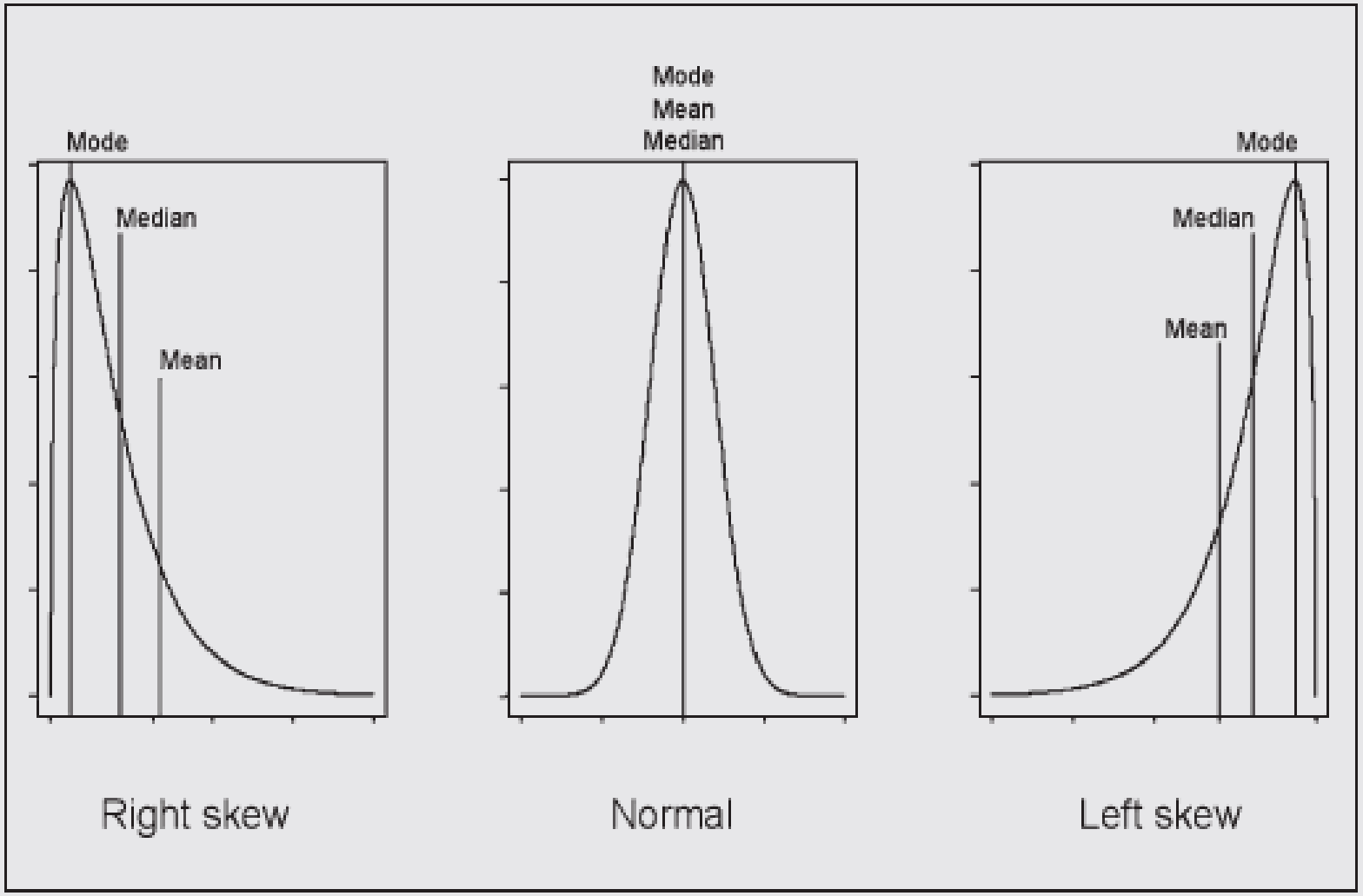
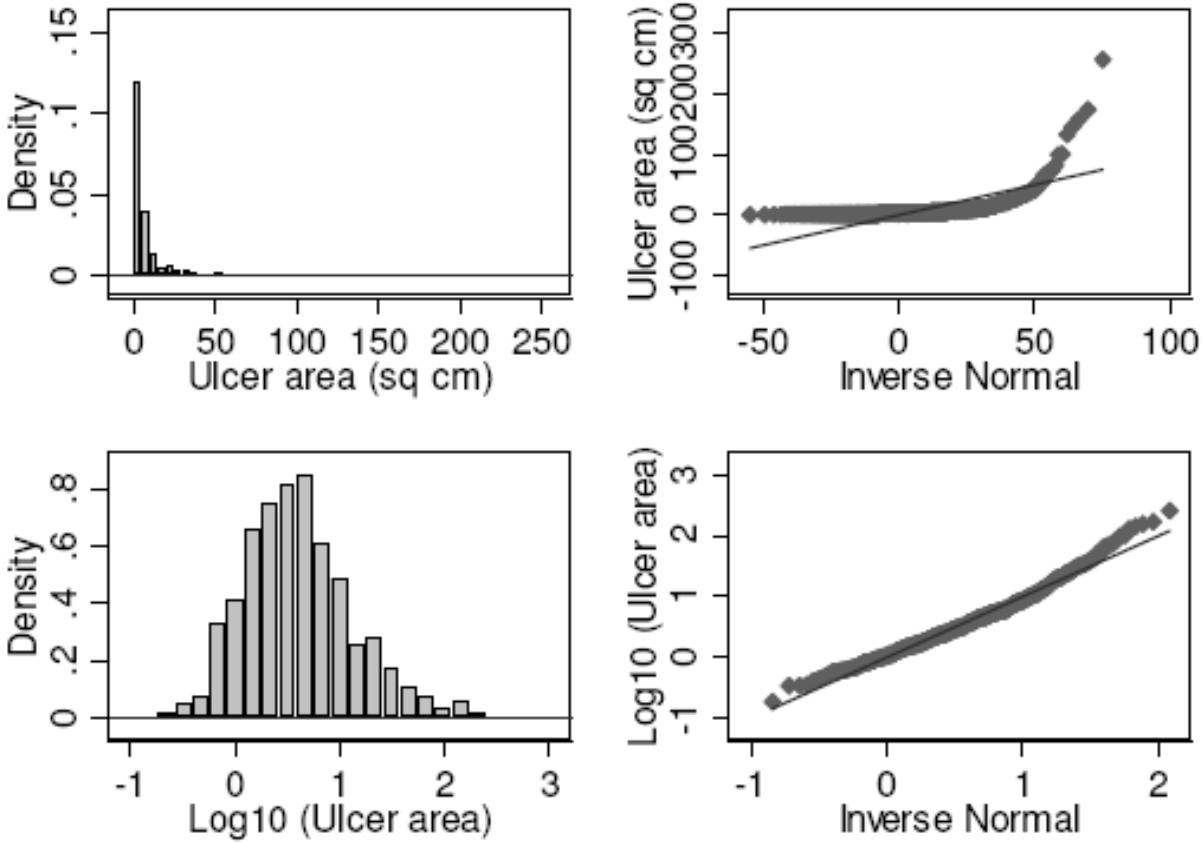


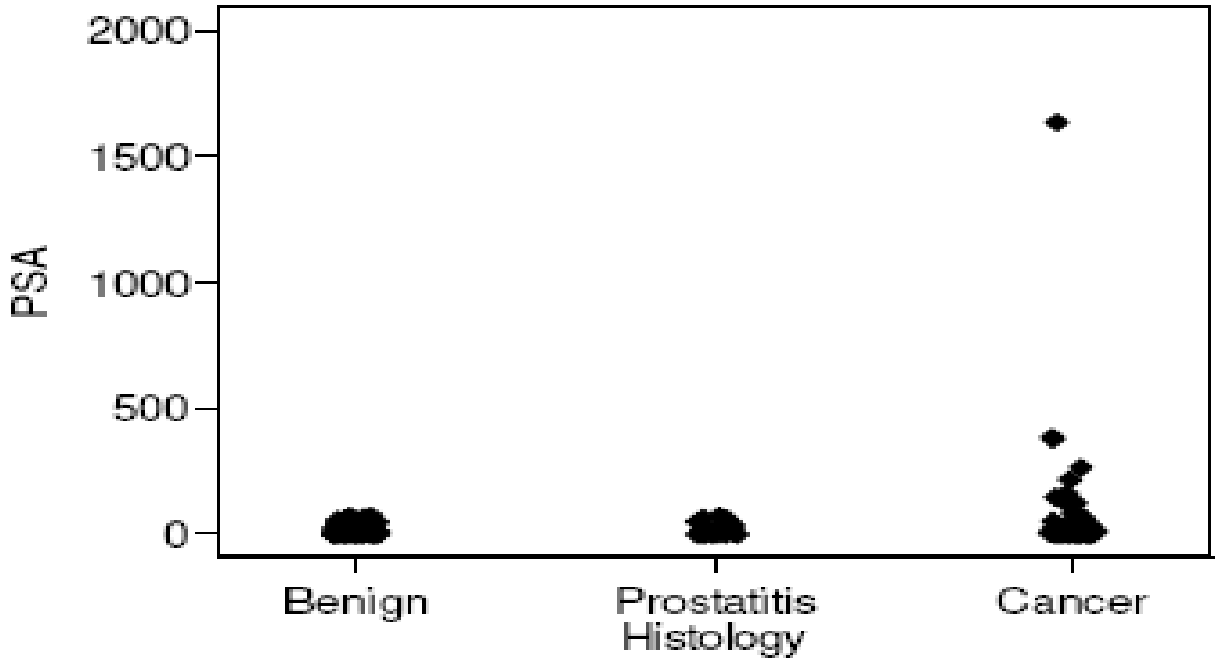
Fig. 2 Distributions of Quantitative Data.

The following figure shows a histogram and Normal plot for the area of venous ulcer at recruitment



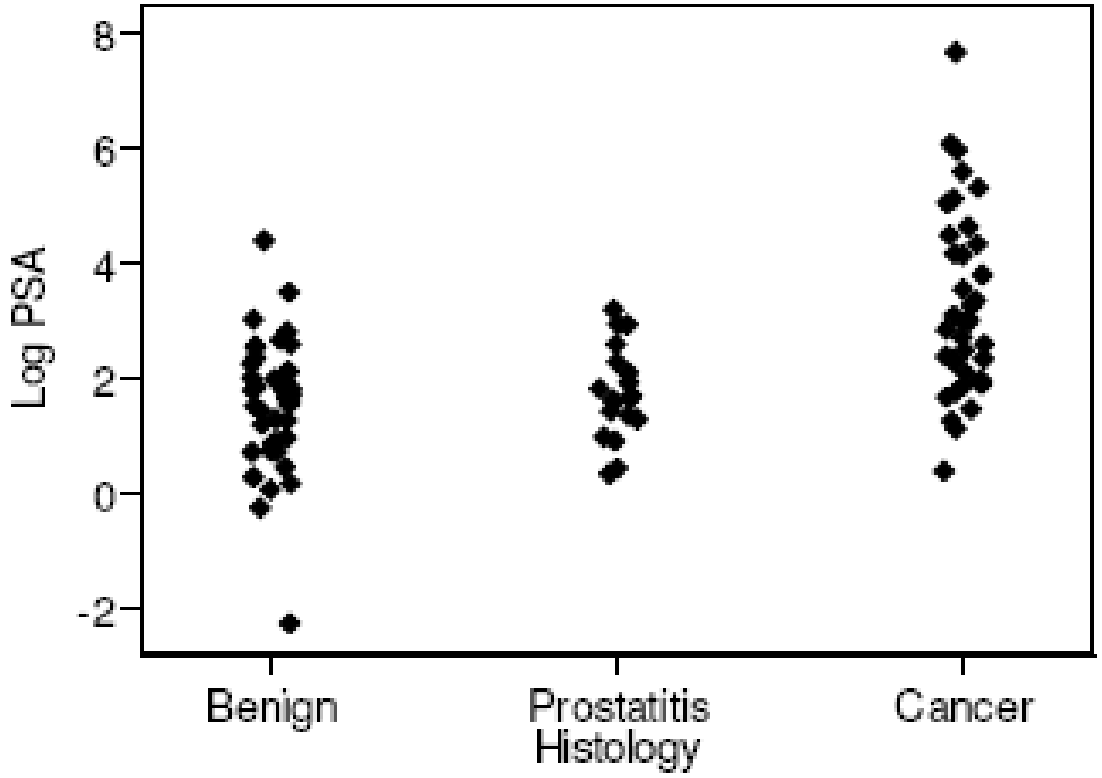
Making a distribution more like the Normal is not the only reason for using a transformation

The following figure shows prostate specific antigen (PSA) for three groups of prostate patients: with benign conditions, with prostatitis, and with prostate cancer



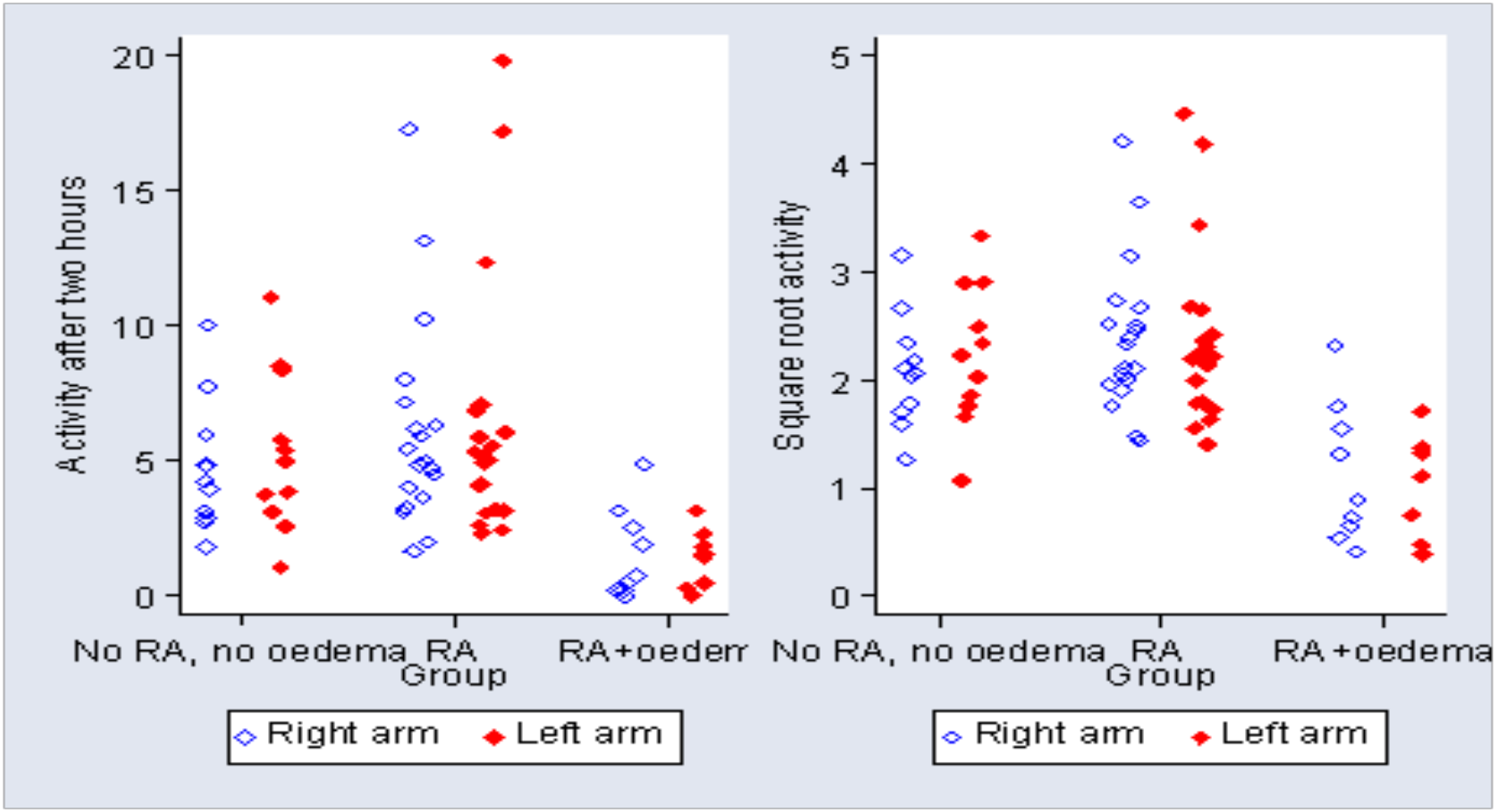
A log transformation of the PSA gives a much clearer picture .

The variability is now much more similar in the three groups



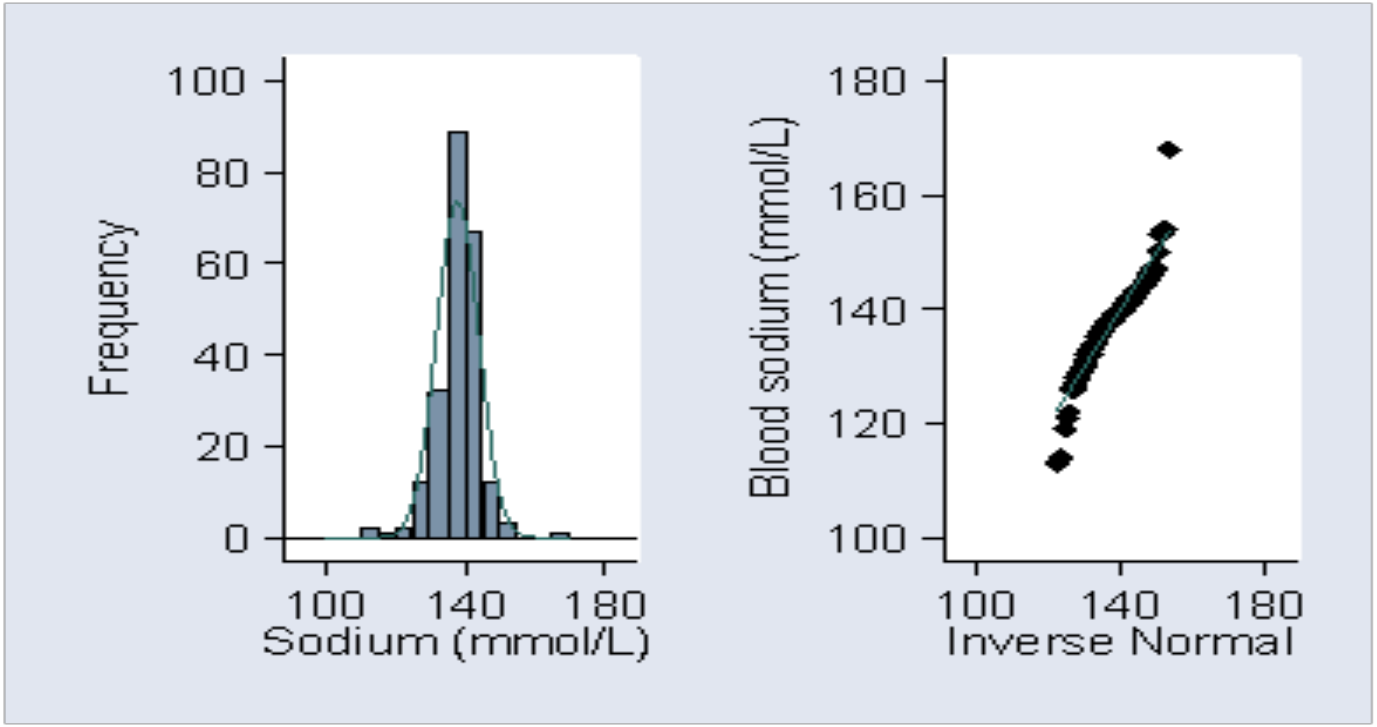
Arm lymphatic flow in rheumatoid arthritis with oedema

•



For example the distribution of blood sodium in ITU patients

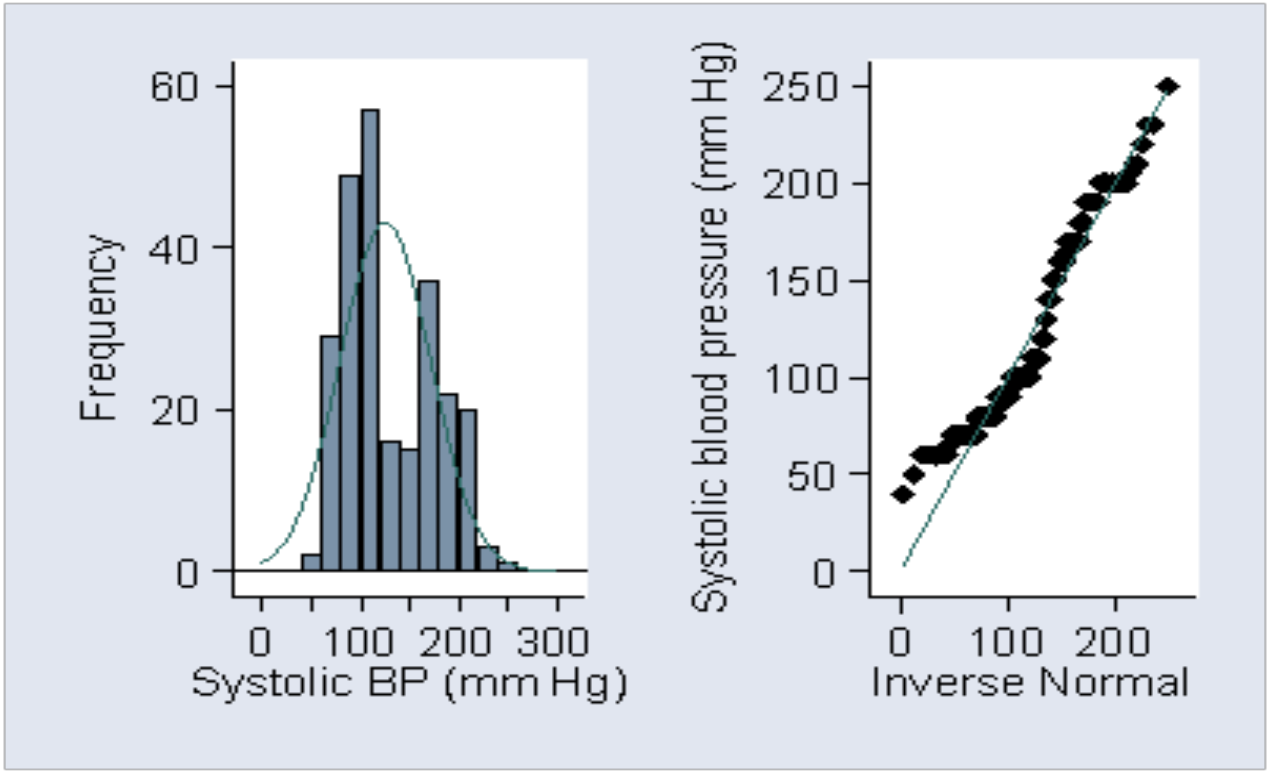
- This is fairly symmetrical, but has longer tails than a Normal distribution .The shape of the Normal plot is first convex then concave



2-Sometimes we have a bimodal distribution, which makes transformation by log, square root or reciprocal ineffective

For example systolic blood pressure in a sample of ITU

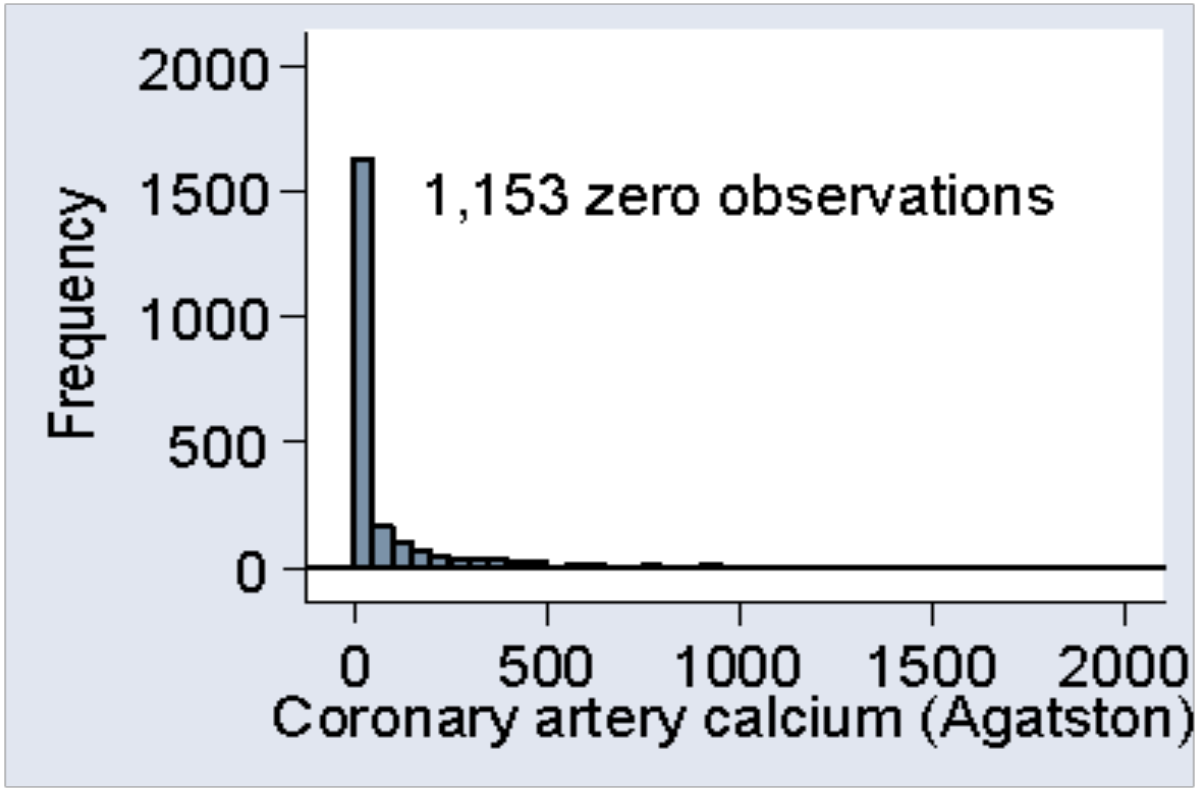
- patients



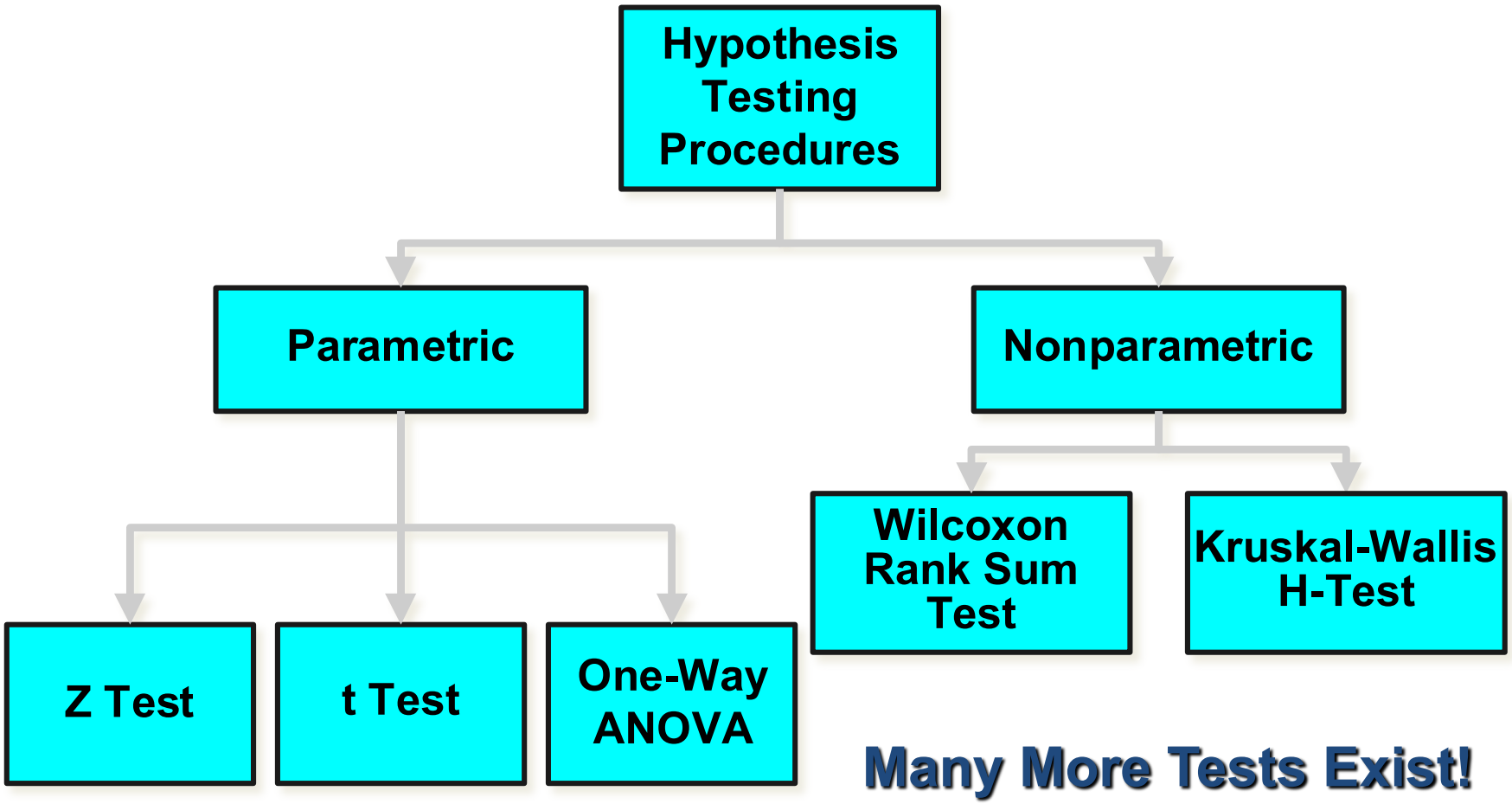
For example the distribution of coronary artery calcium in a large group of patients

More than half of these observations were equal at zero .Any transformation would leave half the observations with the same value, at the extreme of the distribution .It is impossible to transform these data to a Normal distribution

.



Hypothesis Testing Procedures



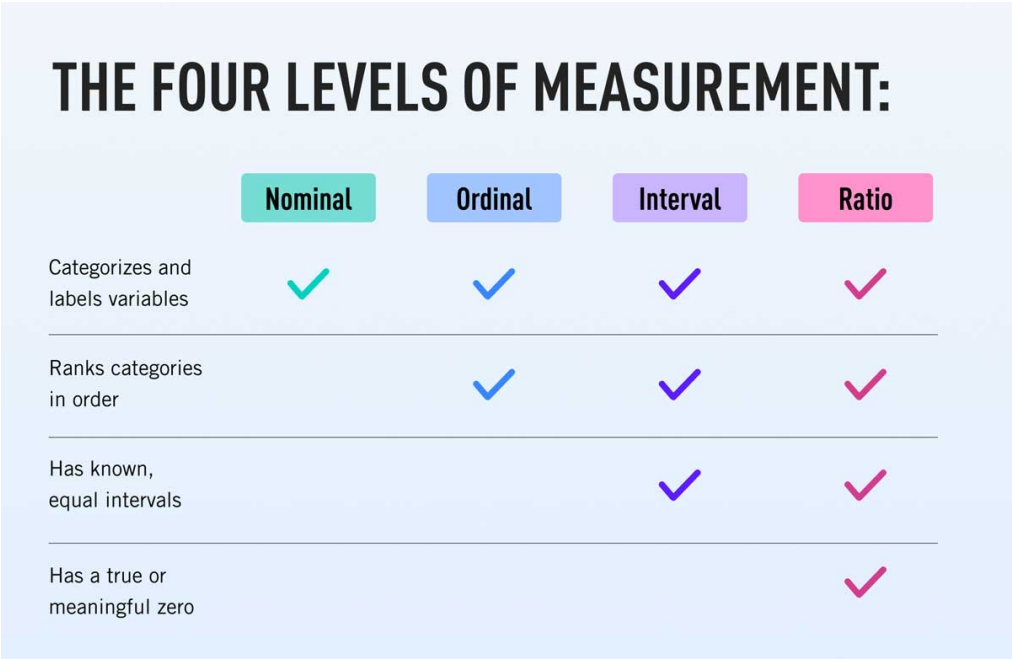
Why non-parametric statistics?

-Need to analyse ‘Crude’ data (nominal, -ordinal)

-Data derived from small samples

-Data that do not follow a normal distribution

-Data of unknown distribution



Nonparametric Test Procedures

A **nonparametric test** is a hypothesis test that does not require any specific conditions about the shape of the populations or the value of any population parameters.

Tests are often called "distribution free" tests.

Disadvantages of Nonparametric Tests

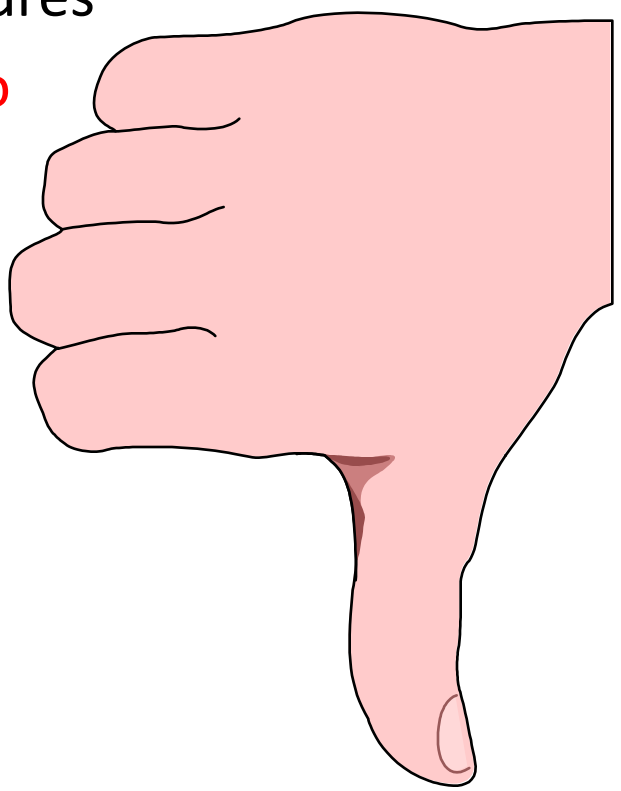
1-May Waste Information

If Data Permit Using Parametric Procedures

Example: Converting Data From Ratio to Ordinal Scale

2-Difficult to Compute by Hand Large Samples

3-Tables Not Widely Available



Advantages of Nonparametric Tests

. 1-Used With All Scales

2-Easier to Compute.

3- Make Fewer Assumptions.

4- Suitable for small sample size.

5-Analysis involves outlier values.

6- No need for population Parameters.

7-Results May Be as Exact as Parametric Procedures