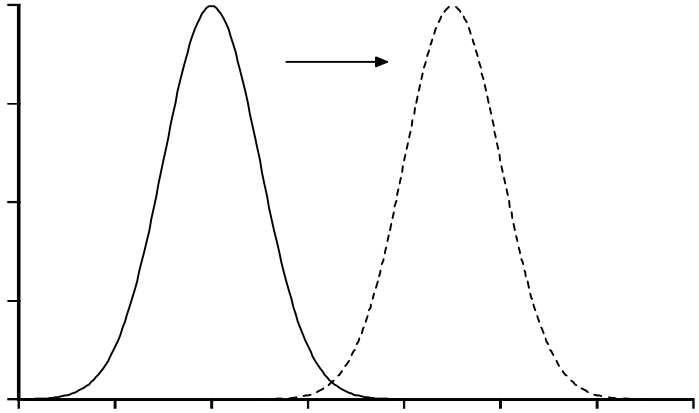
Skewed Data and Non-parametric Methods

Comparing two groups: *t*-test *assumes* data are:

1. Normally distributed,

   and

2. both samples have the same SD
   (i.e. one sample is simply shifted relative to the other)



Violation of Assumptions

1. Ascertain if assumptions hold. Things to bear in mind are:

Slight violations are probably unimportant

Violations of unequal variance worse than violations of Normality

What is an important violation is judged by experience

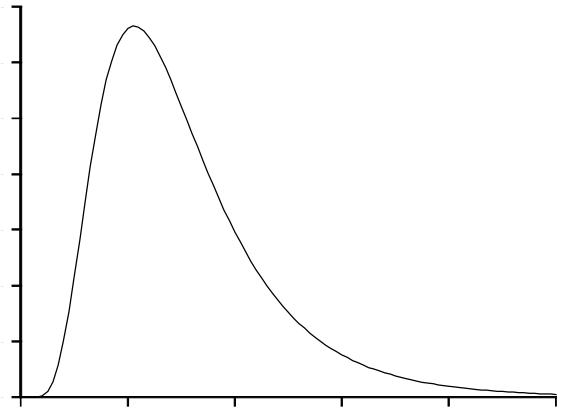
2. Use a method that does not make the assumptions

or

3. Can try to transform data so that assumptions are satisfied

Skewness

Common way data violate assumptions is that their distribution is skewed



The data have asymmetric distribution, with > 50% of population above mode.

Often occurs with measurement that must be positive and SD is large compared with mean.

E.g. if mean-SD <0, for positive variable, Normality cannot be right as it would imply >16% population had negative values.

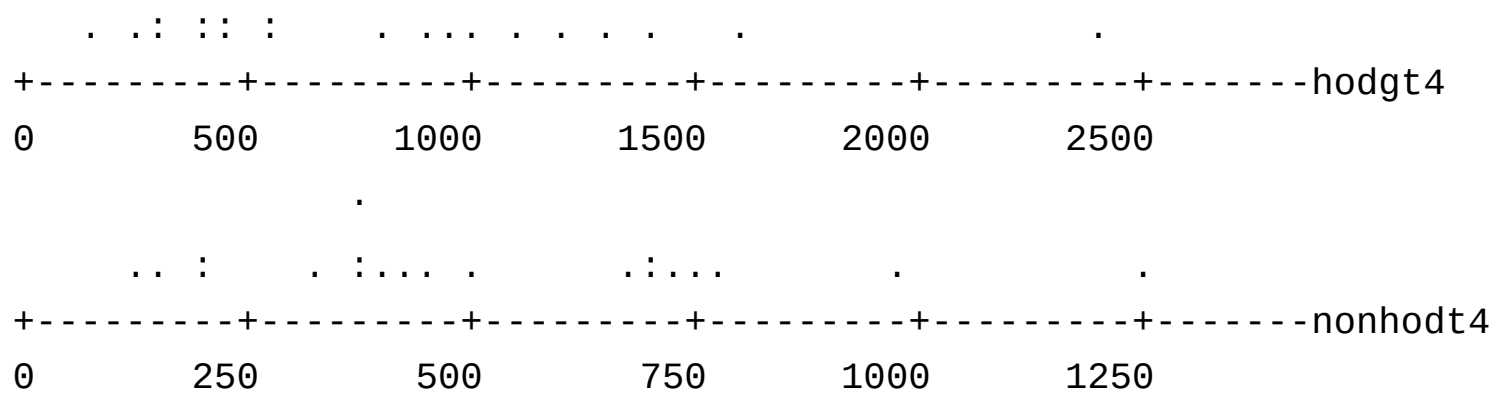
Example: Assessing Assumptions

Compare 20 patients in remission from Hodgkin's disease and 20 patients in remission from non-Hodgkin's disease.

Variable is the number of T_4 cells per mm^3**.** (Altman, DG, *Practical Statistics for Medical Research,* 1991, Chapman & Hall, section 9.7, citing Shapiro *et al.* 1986, *Am J Med Sci*, 293-366-70).

Plotting data is first step: Dotplots (found under Character Graphs under Graph) can be very useful:

```
MTB > DotPlot 'hodgt4' 'nonhodt4'.
```



Assessment of above plots is very subjective.

Additional assessment, slightly less informal, is to look at mean and SD:

```
MTB > Describe  'hodgt4' 'nonhodt4'.

               N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
hodgt4        20      823      682      771      566      127
nonhodt4      20    522.0    433.0    504.1    293.0     65.5
```

Mean is less than 2 x SD for non-Hodgkin's cases and less than 1.5 SD for Hodgkin's cases, so Normality not a good model.

Also, SD is larger when mean is larger

Should avoid using a *t*-test on these data.

Distribution-Free Methods

(aka Non-parametric methods)

The approach here is to use a method that does not assume Normality

For two (unpaired) samples the procedure is the Mann-Whitney test (equivalent to Wilcoxon rank sum test)

Choose Nonparametrics item under Stat menu and select Mann-Whitney…

Other non-parametric procedures exist for other situations, e.g. Wilcoxon signed ranks test is analogue of paired *t*-test

How does rank sum test work?

For illustration, suppose:

Hodgkin's sample contained 3 cases 1378, 958, 431

non-Hodgkin's sample 4 cases 979, 1252, 116, 377

Combine two samples and rank them. The ranks will be 1,2,3,4,…,7 (in general 1,2,3,…,n_1+n_2).

Null hypothesis is that both samples come from the same (arbitrary) distribution.

If true, there will be no tendency for ranks from one sample to be larger than those from other sample

Illustration (continued)

Value	116	377	431	958	979	1252	1378
Rank	1	2	3	4	5	6	7

(Hodgkin's sample underlined)

Once ranks have been found, original data can be discarded for purposes of test.

(This is the way that approach avoids assuming a specific distribution)

Test statistic is sum of ranks from one of the samples, e.g.

1+2+5+6 = 14

Under null hypothesis, expected value is $\frac{4}{7}(1+2+3+4+5+6+7)= 16$

Can also calculate how much observed value is expected to deviate from its expected value

(in general, $\frac{n_1}{n_1+n_2}\times\{1+2+\ldots+(n_1+n_2)\}=\frac{1}{2}n_1(n_1+n_2+1))$

Using the Mann-Whitney test

Output from Minitab as follows

```
MTB > Mann-Whitney 95.0 'hodgt4' 'nonhodt4';
SUBC>   Alternative 0.

Mann-Whitney Confidence Interval and Test

hodgt4      N =  20     Median =       681.5
nonhodt4    N =  20     Median =       433.0
Point estimate for ETA1-ETA2 is        203.0
95.0 Percent C.I. for ETA1-ETA2 is (-26.9,531.9)
W = 475.0
Test of ETA1 = ETA2  vs.  ETA1 ~= ETA2 is significant at 0.0810   ‡
The test is significant at 0.0810 (adjusted for ties)

Cannot reject at alpha = 0.05
```

Mann-Whitney test is show in red (and marked ‡)

For skewed data, mean and SD may not be appropriate, so Median and inter-quartile ranges etc. are used to give estimates of treatment effect.

In Minitab output ETA1, ETA2, are used to refer to medians of the two samples

Point Estimate & Confidence Interval

Above output contains line

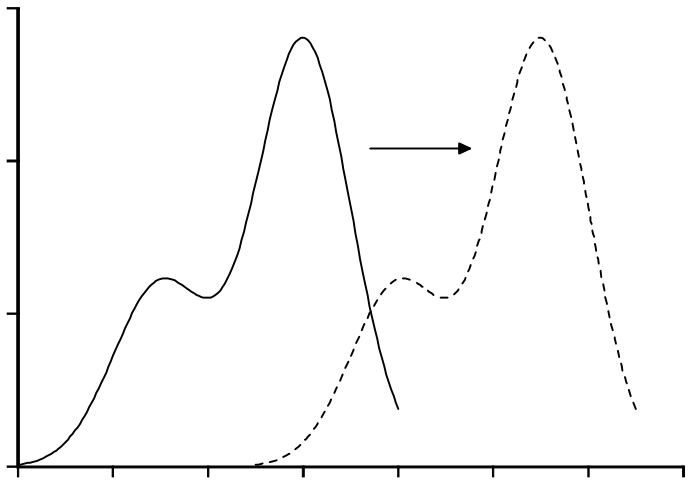
```
hodgt4     N =  20     Median =       681.5
nonhodt4   N =  20     Median =       433.0
Point estimate for ETA1-ETA2 is       203.0
```

Note: 681.5-433.0≠203.0: value of 203.0 is median of all possible differences x-y, where x is one sample and y is in the other.

Confidence interval assumes two distributions have the same shape, which is often not reasonable:

distribution-free ≠ assumption-free

Also, analysis based on ranks must lose information



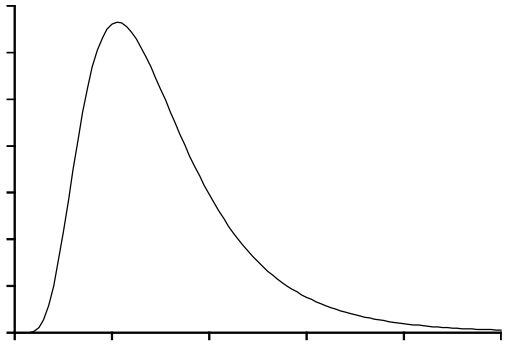
Alternative Approach

If assumptions of *t*-test violated, transform data so that *t*-test can be applied to transformed data.

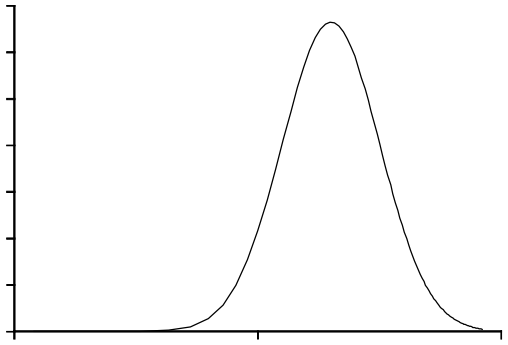
Taking logs of the data is often useful for data that are >0 because:

1. It can get rid of skewness

Before log-transformation



After log-transformation



2. It can turn multiplicative effects into additive ones

Multiplicative into Additive?

Suppose the outcome in control group is Normally distributed

Suppose the effect of treatment is to double the outcome

Then outcome in the control group is also normal

Control	Treated
Mean = μ	Mean = 2μ
SD = σ	SD = 2σ

So a *t*-test cannot be applied because the SDs are unequal

If $X \rightarrow 2X$, then $\log(X) \rightarrow \log(2)+\log(X)$. Mean of log X increases but SD does not

Taking logs

Doesn’t always work, but often helps

Procedure is:

1. Take logs of original data
2. Apply all statistical methods, i.e. calculating means and SDs as well as performing *t*-test, on logged data
3. Transform back to original scale. Care needed to get interpretation correct, do not back transform SDs, just means and confidence intervals

Logging Hodgkin's data

Results, (using logs to base 10) (omitting inessentials)

```
MTB > Describe 'loghod' 'lognhod'.

               N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
loghod        20   2.8172   2.8274   2.8183   0.3076   0.0688
lognhod       20   2.6443   2.6364   2.6513   0.2743   0.0613

MTB > TwoSample 95.0 'loghod' 'lognhod';
SUBC>   Alternative 0;
SUBC>   Pooled.

TWOSAMPLE T FOR loghod VS lognhod
          N      MEAN     STDEV   SE MEAN
loghod   20     2.817     0.308     0.069
lognhod  20     2.644     0.274     0.061

95 PCT CI FOR MU loghod - MU lognhod: ( -0.014,  0.360)

TTEST MU loghod = MU lognhod (VS NE): T= 1.88  P=0.068  DF=  38

POOLED STDEV =      0.291
```

Things to note

1. SDs of two samples of logged data much closer than for unlogged data

2. Visual checks of Normality (e.g. Normal plot) shows logged data more Normal than unlogged

3. P value is 0.068: this is the value to quote, no transformation needed

4. Difference of logged means is 2.817-2.644 = 0.173
   = log (“mean” Hodgkin’s T4 counts)-log (“mean” non-Hodgkin’s T4 counts)
   $= \log\left(\frac{\text{"mean" Hodgkin's T4 counts}}{\text{"mean" non-Hodgkin's T4 counts}}\right)$

Hence $\frac{\text{"mean" Hodgkin's T4 counts}}{\text{"mean" non-Hodgkin's T4 counts}}$ = antilog(0.173)=1.49

So “mean” T4 count is about 1.5 times bigger in Hodgkin’s than non-Hodgkin’s

95% of the time the multiplying factor will be between antilog(-0.014) and antilog(0.360) (from confidence interval), that is, between 0.97, 2.29

"Mean"

In above mean is written in quotes to signify it does not mean the usual (arithmetic) mean, but the geometric mean, a measure which is more suited to skewed data.

Consider data 2, 3, 5 or general sample $x_1, \ldots, x_n$

	Example	General case
Arithmetic mean (AM)	$\frac{2+3+5}{3} = 3.333$	$\frac{x_1+\ldots+x_n}{n}$
Geometric mean (GM)	$(2\times3\times5)^{1/3} = 3.107$	$(x_1 \times x_2 \ldots \times x_n)^{1/n}$

GM of unlogged data= antilog of AM of logged data

Further reading: Bland, chapter 12 (2nd ed.); Altman, section 9.6; articles by Bland & Altman in *BMJ*, vol. 312, 1996: p700, p. 770, p.1153