STAT 401A - Statistical Methods for Research Workers

Modeling assumptions

Jarad Niemi (Dr. J)

Iowa State University

last updated: September 15, 2014

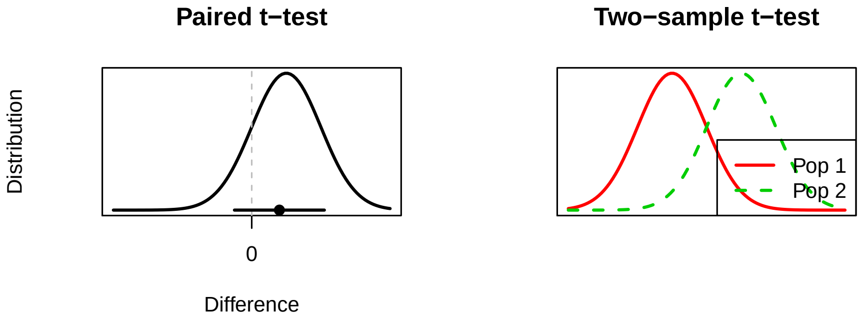
Normality assumptions

In the paired t-test, we assume

$$D_i \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

In the two-sample t-test, we assume

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2).$$



Normality assumptions

In the paired t-test, we assume

$$D_i \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

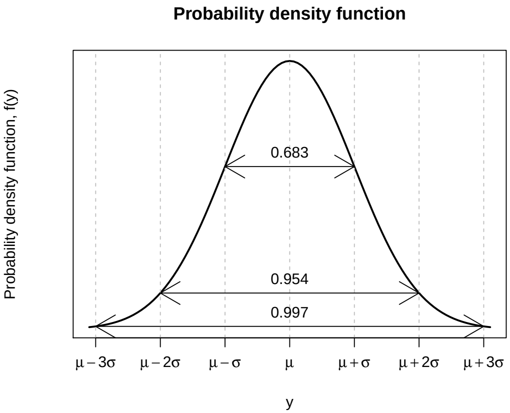
In the two-sample t-test, we assume

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2).$$

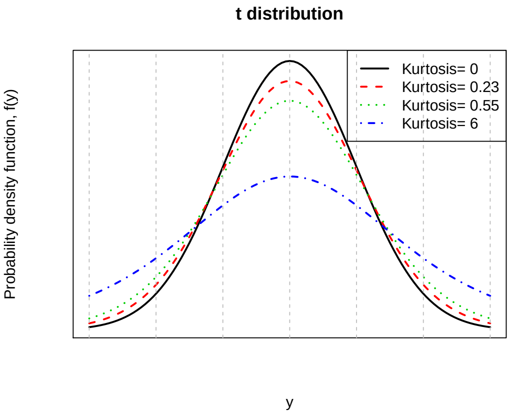
Key features of the normal distribution assumption:

- Centered at the mean (expectation) μ
- Standard deviation describes the spread
- Symmetric around μ (no skewness)
- Non-heavy tails, i.e. outliers are rare (no kurtosis)

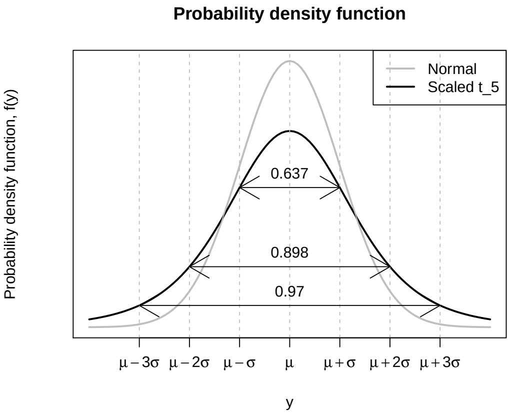
Normality assumptions



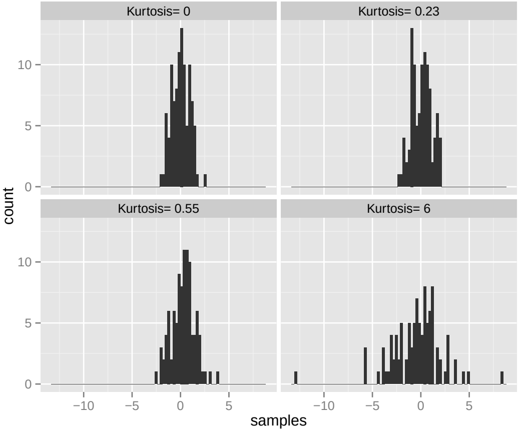
Kurtosis (heavy-tailedness)



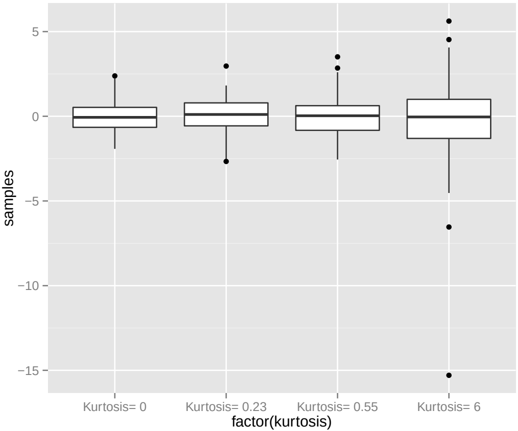
Kurtosis (heavy-tailedness)



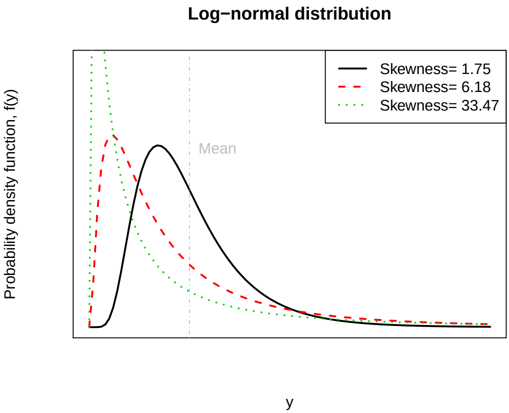
Kurtosis (heavy-tailedness)



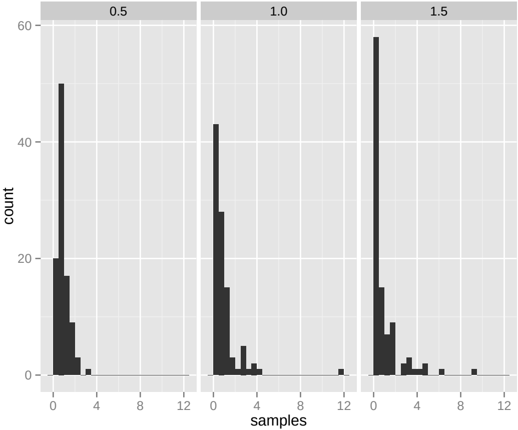
Kurtosis (heavy-tailedness)



Skewness


Log−normal distribution
Skewness= 1.75
Skewness= 6.18
Skewness= 33.47
Mean
Probability density function, f(y)
y

Samples from skewed distributions



Robustness

Definition

A statistical procedure is robust to departures from a particular assumption if it is valid even when the assumption is not met.

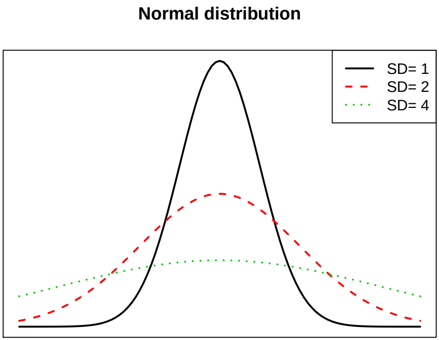
Remark If a 95% confidence interval is robust to departures from a particular assumption, the confidence interval should cover the true value about 95% of the time.

Robustness to skewness and kurtosis

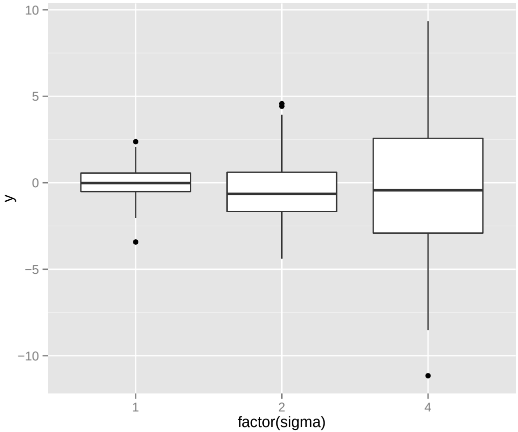
Percentage of 95% confidence intervals that cover the true difference in means in an equal-sample two-sample t-test with non-normal populations (where the distributions are the same other than their means).

sample size	strongly skewed	moderately skewed	mildly skewed	heavy-tailed	short-tailed
5	95.5	95.4	95.2	98.3	94.5
10	95.5	95.4	95.2	98.3	94.6
25	95.3	95.3	95.1	98.2	94.9
50	95.1	95.3	95.1	98.1	95.2
100	94.8	95.3	95.0	98.0	95.6

Differences in variances



Differences in variances



Robustness to differences in variances

Percentage of 95% confidence intervals that cover the true difference in means in an equal-sample two-sample t-test ($r = \sigma_1/\sigma_2$).

n1	n2	r=1/4	r=1/2	r=1	r=2	r=4
10	10	95.2	94.2	94.7	95.2	94.5
10	20	83.0	89.3	94.4	98.7	99.1
10	40	71.0	82.6	95.2	99.5	99.9
100	100	94.8	96.2	95.4	95.3	95.1
100	200	86.5	88.3	94.8	98.8	99.4
100	400	71.6	81.5	95.0	99.5	99.9

Outliers

Definition

A statistical procedure is resistant if it does not change very much when a small part of the data changes, perhaps drastically.

Identify outliers:

1. If recording errors, fix.
2. If outlier comes from a different population, remove and report.
3. If results are the same with and without outliers, report with outliers.
4. If results are different, use resistant analysis or report both analyses.

Common ways for independence to be violated

- Cluster effect
  - e.g. pigs in a pen
- Serial effect
  - e.g. measurements in time with drifting scale
- Spatial effect
  - e.g. corn yield plots (drainage)

Common transformations for data

From: http://en.wikipedia.org/wiki/Data_transformation_(statistics)

Definition

In statistics, data transformation refers to the application of a deterministic mathematical function to each point in a data set that is, each data point y_i is replaced with the transformed value $z_i = f(y_i)$, where f is a function.

The most common transformations are

- If y is a proportion, then $f(y) = sin^{-1}(\sqrt{y})$.
- If y is a count, then $f(y) = \sqrt{y}$.
- If y is positive and right-skewed, then $f(y) = \log(y)$, the *natural logarithm* of y.

Remark Since $\log(0) = -\infty$, the logarithm cannot be used directly when some y_i are zero. In these cases, use $\log(y + c)$ where c is something small relative to your data, e.g. half of the minimum non-zero value.

Log transformation

Consider two-sample data and let $z_{ij} = log(y_{ij})$. Now, run a two-sample t-test on the z's. Then we assume

$$Z_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

and the quantity $\overline{Z}_2 - \overline{Z}_1$ estimates the "difference in population means on the (natural) log scale". The quantity $\exp\left(\overline{Z}_2 - \overline{Z}_1\right) = e^{\overline{Z}_2 - \overline{Z}_1}$ estimates

$$\frac{\text{Median of population 2}}{\text{Median of population 1}}$$

on the original scale or, equivalently, it estimates the multiplicative effect of moving from population 1 to population 2.

Log transformation interpretation

If we have a randomized experiment:

Remark It is estimated that the response of an experimental unit to treatment 2 will be $\exp\left(\overline{Z}_2 - \overline{Z}_1\right)$ times as large as its response to treatment 1.

If we have an observational study:

Remark It is estimated that the median for population 2 is $\exp\left(\overline{Z}_2 - \overline{Z}_1\right)$ times as large as the median for population 1.

Confidence intervals with log transformation

If $z_{ij} = log(y_{ij})$ and we assume

$$Z_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2),$$

then a $100(1-\alpha)\%$ two-sided confidence interval for $\mu_2 - \mu_1$ is

$$(L, U) = \overline{Z}_2 - \overline{Z}_1 \pm t_{n_1+n_2-2}(1-\alpha/2) SE\left(\overline{Z}_2 - \overline{Z}_1\right).$$

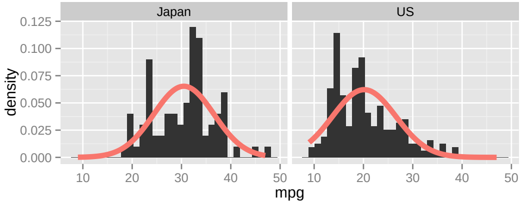
A $100(1-\alpha)\%$ confidence interval for

$$\frac{\text{Median of population 2}}{\text{Median of population 1}}$$

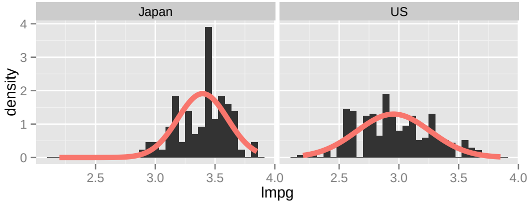
is (e^L, e^U).

Miles per gallon data

Untransformed:

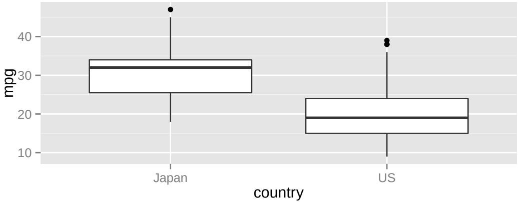


Logged:

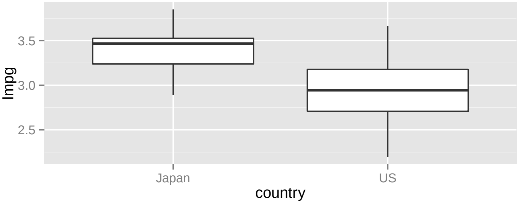


Miles per gallon data

Untransformed:



Logged:



Equal variances?

We might also be concerned about the assumption of equal variances.

Untransformed:

country	n	mean	sd
Japan	79	30.48	6.11
US	249	20.14	6.41

the ratio of sample standard deviations is around 1.05 and there are 3 times as many observations in the US.

Logged:

country	n	mean	sd
Japan	79	3.40	0.21
US	249	2.96	0.31

Now the ratio of standard deviations is 1.5 which argues for not using the logarithm.

95% two-sample CI for the ratio by hand

country	n	mean	sd
Japan	79	3.40	0.21
US	249	2.96	0.31

Choose group 2 to be Japan and group 1 to be the US:

$$\begin{aligned}
\alpha &= 0.05 \\
n_1 + n_2 - 2 &= 249 + 79 - 2 = 326 \\
t_{n_1+n_2-2}(1-\alpha/2) &= t_{326}(0.975) = 1.96 \\
\overline{Z}_2 - \overline{Z}_1 &= 3.40 - 2.96 = 0.44 \\
s_p &= \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(249-1)0.31^2 + (79-1)0.21^2}{249+79-2}} = 0.29 \\
SE\left(\overline{Z}_2 - \overline{Z}_1\right) &= s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.29\sqrt{\frac{1}{249} + \frac{1}{79}} = 0.037
\end{aligned}$$

Thus a 95% two-sided confidence interval for the difference (on the log scale) is

$$\begin{aligned}
(L, U) &= \overline{Z}_2 - \overline{Z}_1 \pm t_{n_1+n_2-2}(1-\alpha/2) SE\left(\overline{Z}_2 - \overline{Z}_1\right) \\
&= 0.44 \pm 1.96 \times 0.037 \\
&= (0.37, 0.51)
\end{aligned}$$

and a 95% two-sided confidence interval for the ratio (on the original scale) is

$$\left(e^L, e^U\right) = \left(e^{0.37}, e^{0.51}\right) = (1.45, 1, 67)$$

Using R for t-test using logarithms

```
t = t.test(log(mpg)~country, d, var.equal=TRUE)
t$estimate # On log scale

mean in group Japan    mean in group US
              3.396               2.955

exp(t$estimate) # On original scale

mean in group Japan    mean in group US
              29.85               19.21

exp(t$estimate[1]-t$estimate[2]) # Ratio of medians (Japan/US)

mean in group Japan
              1.554

exp(t$conf.int) # Confidence interval for ratio of medians

[1] 1.445 1.672
attr(,"conf.level")
[1] 0.95
```

SAS code for t-test using logarithms

```
DATA mpg;
  INFILE 'mpg.csv' DELIMITER=',' FIRSTOBS=2;
INPUT mpg country $;

PROC TTEST DATA=mpg TEST=ratio;
CLASS country;
VAR mpg;
run;
```

SAS output for t-test using logarithms

```
                          The TTEST Procedure

                           Variable:  mpg

                         Geometric    Coefficient
       country         N      Mean   of Variation    Minimum    Maximum

       Japan          79   29.8525         0.2111    18.0000    47.0000
       US            249   19.2051         0.3147     9.0000    39.0000
       Ratio (1/2)          1.5544         0.2928

                    Geometric                        Coefficient
country      Method          Mean       95% CL Mean     of Variation       95% CL CV

Japan                     29.8525   28.4887  31.2817          0.2111    0.1820   0.2514
US                        19.2051   18.4825  19.9560          0.3147    0.2882   0.3467
Ratio (1/2)  Pooled        1.5544    1.4452   1.6719          0.2928    0.2712   0.3183
Ratio (1/2)  Satterthwaite 1.5544    1.4636   1.6508

                                Coefficients
         Method          of Variation        DF    t Value    Pr > |t|

         Pooled          Equal              326      11.91      <.0001
         Satterthwaite   Unequal         193.33      14.46      <.0001

                         Equality of Variances

             Method       Num DF    Den DF    F Value    Pr > F

             Folded F        248        78       2.17    0.0001
```

Conclusion

Japanese median miles per gallon is 1.55 [95% CI (1.46,1.65)] times as large as US median miles per gallon.

OR

Japenese median miles per gallon is 55% [95% CI (46%,65%)] larger than US median miles per gallon.

Unequal standard deviations

The two-sample t-test tools assume either

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2) \qquad \text{or} \qquad Z_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$$

depending on whether we were working on the original scale (Y) or log scale (Z), respectively.

But what if we don't believe the variances in the two populations are equal, e.g. in the log transformed miles per gallon data set?

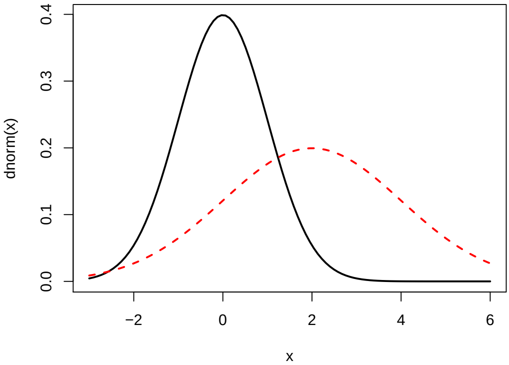
Instead compare

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma_j^2) \qquad \text{or} \qquad Z_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma_j^2),$$

i.e. the populations have unequal variances. But still test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ or construct a confidence interval for $\mu_2 - \mu_1$.

Visualization of two normals with unequal standard deviations

```
curve(dnorm, -3, 6, lwd=2)
curve(dnorm(x, 2, 2), lwd=2, col=2, lty=2, add=TRUE)
```



Welch's SE with Satterthwaite's approximation to df

Estimate of $(\mu_2 - \mu_1)$:

$$\overline{Y}_2 - \overline{Y}_1$$

Standard error:

$$SE_W\left(\overline{Y}_2 - \overline{Y}_1\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degrees of freedom using the Satterthwaite's approximation:

$$df_W = \frac{SE_W\left(\overline{Y}_2 - \overline{Y}_1\right)^4}{\frac{SE\left(\overline{Y}_2\right)^4}{n_2-1} + \frac{SE\left(\overline{Y}_1\right)^4}{n_1-1}}$$

where

$$SE\left(\overline{Y}_2\right) = \frac{s_2}{\sqrt{n_2}} \qquad \text{and} \qquad SE\left(\overline{Y}_1\right) = \frac{s_1}{\sqrt{n_1}}$$

(which is the same formula as in the paired t-test)

Welch's t-test and CI

Welch's t-test has test statistic:

$$t = \frac{\text{(Estimate-Parameter)}}{\text{SE(Estimate)}} = \frac{\overline{Y}_2 - \overline{Y}_1 - (\mu_2 - \mu_1)}{SE_W\left(\overline{Y}_2 - \overline{Y}_1\right)}$$

which has a t distribution with (approximately) df_W degrees of freedom if the null hypothesis is true. Calculate the pvalue

- Two-sided ($H_1 : \mu_2 \neq \mu_1$): $p = 2P(t_{df_W} < -|t|)$
- One-sided ($H_1 : \mu_2 > \mu_1$): $p = P(t_{df_W} < -t)$
- One-sided ($H_1 : \mu_2 < \mu_1$): $p = P(t_{df_W} < t)$

Two-sided $100(1-\alpha)\%$ confidence interval for $\mu_2 - \mu_1$:

$$\overline{Y}_2 - \overline{Y}_1 \pm t_{df_W}(1-\alpha/2) SE_W\left(\overline{Y}_2 - \overline{Y}_1\right)$$

Are the variances equal?

Suppose

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma_j^2)$$

and you want to test $H_0 : \sigma_1 = \sigma_2$ vs $H_1 : \sigma_1 \neq \sigma_2$.

You can use an F-test and its associated pvalue. If the pvalue is small, e.g. less than 0.05, then we reject H_0. If the pvalue is not small, then we fail to reject H_0, but this does not mean the variances are not equal.

(Section 4.5.3) discusses another approach called Levene's test

Welch's test and CI using R

```
var.test(mpg~country,d) # F-test



F test to compare two variances

data:  mpg by country
F = 0.9066, num df = 78, denom df = 248, p-value = 0.6194
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6423 1.3246
sample estimates:
ratio of variances
            0.9066


(t=t.test(mpg~country, d, var.equal=FALSE))



Welch Two Sample t-test

data:  mpg by country
t = 12.95, df = 136.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.758 11.915
sample estimates:
mean in group Japan    mean in group US
              30.48               20.14
```

SAS code for two-sample t-test

```
DATA mpg;
    INFILE 'mpg.csv' DELIMITER=',' FIRSTOBS=2;
    INPUT mpg country $;

PROC TTEST DATA=mpg;
    CLASS country;
    VAR mpg;
    RUN;
```

SAS output for t-test

```
                          The TTEST Procedure

                           Variable:  mpg

     country          N        Mean     Std Dev     Std Err     Minimum     Maximum

     Japan           79     30.4810      6.1077      0.6872     18.0000     47.0000
     US             249     20.1446      6.4147      0.4065      9.0000     39.0000
     Diff (1-2)             10.3364      6.3426      0.8190

country       Method              Mean      95% CL Mean         Std Dev      95% CL Std Dev

Japan                          30.4810    29.1130  31.8491       6.1077      5.2814   7.2429
US                             20.1446    19.3439  20.9452       6.4147      5.8964   7.0336
Diff (1-2)    Pooled           10.3364     8.7252  11.9477       6.3426      5.8909   6.8699
Diff (1-2)    Satterthwaite    10.3364     8.7576  11.9152

               Method           Variances        df    t Value    Pr > |t|

               Pooled           Equal           326      12.62      <.0001
               Satterthwaite    Unequal      136.87      12.95      <.0001

                         Equality of Variances

                  Method      Num df    Den df    F Value    Pr > F

                  Folded F       248        78       1.10    0.6194
```

```
var.test(log(mpg)~country,d)



F test to compare two variances

data:  log(mpg) by country
F = 0.4617, num df = 78, denom df = 248, p-value = 0.0001055
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3271 0.6745
sample estimates:
ratio of variances
            0.4617


(t = t.test(log(mpg)~country, d, var.equal=FALSE))



Welch Two Sample t-test

data:  log(mpg) by country
t = 14.46, df = 193.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3809 0.5013
sample estimates:
mean in group Japan    mean in group US
              3.396               2.955


exp(t$conf.int)


[1] 1.464 1.651
attr(,"conf.level")
```

SAS code for t-test using logarithms

```
DATA mpg;
  INFILE 'mpg.csv' DELIMITER=',' FIRSTOBS=2;
  INPUT mpg country $;

PROC TTEST DATA=mpg TEST=ratio;
CLASS country;
VAR mpg;
run;
```

SAS output for t-test using logarithms

The TTEST Procedure

Variable: mpg

country	N	Geometric Mean	Coefficient of Variation	Minimum	Maximum
Japan	79	29.8525	0.2111	18.0000	47.0000
US	249	19.2051	0.3147	9.0000	39.0000
Ratio (1/2)		1.5544	0.2928		

country	Method	Geometric Mean	95% CL Mean		Coefficient of Variation	95% CL CV	
Japan		29.8525	28.4887	31.2817	0.2111	0.1820	0.2514
US		19.2051	18.4825	19.9560	0.3147	0.2882	0.3467
Ratio (1/2)	Pooled	1.5544	1.4452	1.6719	0.2928	0.2712	0.3183
Ratio (1/2)	Satterthwaite	1.5544	1.4636	1.6508			

Method	Coefficients of Variation	DF	t Value	Pr > |t|
Pooled	Equal	326	11.91	<.0001
Satterthwaite	Unequal	193.33	14.46	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	248	78	2.17	0.0001

Summary

Two-sample t tools assumptions

- Normality
  - No skewness (take logs?)
  - No heavy tails
- Equal variances
  - Test: F-test or Levene's test
  - Use Welch's two-sample t-test and CI
- Independence (use random effects or avoid)
  - Cluster
  - Serial
  - Spatial