

Power

The *power* of a hypothesis test is the probability of rejecting the null H_0 . A *Type II* error is not rejecting H_0 when H_1 is really true; in this case the power of the test is the probability of rightly rejecting the null. The probability of a Type II error is commonly denoted β .

$$\text{power} = P(\text{reject } H_0 | H_1 \text{ true}) = 1 - P(\text{accept } H_0 | H_1 \text{ true}) = 1 - \beta.$$

Obviously, more power is good. A more powerful test will give smaller p -values, and will reject H_0 more easily. We will discuss power and sample size calculations later.

More on the two-sample problem Section 11.2.1 (pp. 421–425, 428-432).

We've discussed the two-sample normal model with equal variances. Today we'll discuss a generalization to unequal variances and an alternative test known as the Mann-Whitney nonparametric two-sample test. Outline for today:

- A bit on testing for equal variances and normality within groups.
- Two-sample test for normal data with unequal variances.
- Nonparametric two-sample Mann-Whitney test for difference in medians.

Checking for constant variance

A test based on normality in the two samples is `var.test()` for two samples or `bartlett.test()` for multiple samples.

Both of these are referred to as Bartlett's test for homogeneity of variances across populations. The assumptions in the two sample case are

$$Y_{11}, \dots, Y_{1n_1} \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2) \text{ indep. of } Y_{21}, \dots, Y_{2n_2} \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2),$$

and the null hypothesis is $H_0 : \sigma_1 = \sigma_2$.

The data should be in two vectors, the outcomes and a group indicator. The group indicator is typically a factor, but could be numeric as well. This R code tests for equal variances in two groups of sample sizes $n_1 = n_2$ randomly drawn from $N(5, 1)$ and $N(5, 4)$ distributions:

```
> help(var.test)
> help(rnorm)
> y <- c(rnorm(20,mean=5,sd=1),rnorm(20,mean=5,sd=2))
> g <- c(rep(1,20),rep(2,20))
> var.test(y~g)
      F test to compare two variances
data:  y by g
F = 0.3544, num df = 19, denom df = 19, p-value = 0.02892
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1402816 0.8954113
sample estimates: ratio of variances
 0.3544147
```

An alternative, robust test for constant variance across several groups that does not assume normal data is Levene's test. There's at least two packages that have `levene.test()`, `Rcmdr` and `lawstat`. The first also provides a GUI to do basic exploratory statistics as well as two-sample tests and ANOVA. The second package is a good deal smaller in size.

Under **Packages** choose **Install package(s)...**, then pick a mirror if asked, then pick, e.g., `lawstat` and hit **OK**. Then under **Packages** pick **Load package...** and select the package you just downloaded and hit **OK**. Try `help(levene.test)`.

Levene's test does not assume that the two groups have normal population density curves but only assumes for the two samples, Y_{11}, \dots, Y_{1n_1} and Y_{21}, \dots, Y_{2n_2} , that $\text{Var}(Y_{1j}) = \sigma_1^2$ and $\text{Var}(Y_{2j}) = \sigma_2^2$ and tests the null $H_0 : \sigma_1 = \sigma_2$.

For some randomly generated data, the two tests give roughly the same results:

```
> data <- c(rnorm(20,0,1),rnorm(20,0,2))
> group <- c(rep(1,20),rep(2,20))
> var.test(data~group)
      F test to compare two variances
data:  data by group
F = 0.3124, num df = 19, denom df = 19, p-value = 0.01479
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1236705 0.7893833
sample estimates: ratio of variances
 0.3124475
> levene.test(data,group)
      Classical Levene's test based on the absolute deviations from the mean
data:  data Test Statistic = 4.7442, p-value = 0.03567
```

Here the data truly are normal and Bartlett's test, which assumes normality, provides a bit more power to reject the null $H_0 : \sigma_1 = \sigma_2$ (i.e. Bartlett's test has a smaller p -value).

Review of two sample CI and t -test

Main assumption is normal data with equal population variances, i.e. $\sigma_1 = \sigma_2$. Let (n_1, \bar{Y}_1, S_1^2) and (n_2, \bar{Y}_2, S_2^2) be the sample sizes, sample means and sample variances from the two samples.

The standard CI for $\mu_1 - \mu_2$ is given by

$$\text{Lower} = (\bar{Y}_1 - \bar{Y}_2) - t_{crit} \text{se}(\bar{Y}_1 - \bar{Y}_2)$$

$$\text{Upper} = (\bar{Y}_1 - \bar{Y}_2) + t_{crit} \text{se}(\bar{Y}_1 - \bar{Y}_2)$$

The t -statistic for testing $H_0 : \mu_1 - \mu_2 = 0$ ($\mu_1 = \mu_2$) against $H_1 : \mu_1 - \mu_2 \neq 0$ ($\mu_1 \neq \mu_2$) is given by

$$T_0 = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{se}(\bar{Y}_1 - \bar{Y}_2)}.$$

A value of T_0 large in magnitude relative to a $t_{n_1+n_2-2}$ density gives evidence in favor of $H_1 : \mu_1 \neq \mu_2$.

The standard error of $\bar{Y}_1 - \bar{Y}_2$ used in both the CI and the test is given by

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = S_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Where the *pooled variance estimator*,

$$S_{pooled}^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2},$$

is our best unbiased estimate of the common population variance σ^2 . The pooled estimator gives more weight to the larger sample.

A critical value t_{crit} such that we reject when $|T_0| > t_{crit}$ is chosen so that the probability of *wrongly rejecting the null hypothesis* is fixed at α , typically $\alpha = 0.05$. Alternatively, a *p-value* can be computed as

$$p\text{-value} = P(|T_0| > |t_0|),$$

and H_0 rejected if *p-value* is less than α .

The pooled CI and tests are sensitive to the normality and equal standard deviation assumptions. The observed data can be used to assess the reasonableness of these assumptions. You should look at boxplots to assess normality and to assess the assumption $\sigma_1 = \sigma_2$. Formal tests of these assumptions have been discussed.

Satterthwaite's Method

Satterthwaite's method assumes normality, but does not require equal population standard deviations. Satterthwaite's procedures are somewhat conservative, and adjust $se(\bar{Y}_1 - \bar{Y}_2)$ and df to account for unequal population variances. The model has less stringent assumptions:

$$Y_{11}, \dots, Y_{1n_1} \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2) \text{ independent of } Y_{21}, \dots, Y_{2n_2} \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2).$$

There are four unknown population parameters in the model $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. Satterthwaite's method uses the same CI and test statistic formula, with a modified standard error:

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}},$$

and degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Note that $df = n_1 + n_2 - 2$ when $n_1 = n_2$ and $s_1 = s_2$. The Satterthwaite and pooled variance procedures usually give similar results when $s_1 \approx s_2$.

Example: Androstenedione Levels in Diabetics

The data are from independent samples of diabetic men and women. For each individual, the androstenedione level (a hormone) was recorded. Let μ_m = mean androstenedione level for the population of diabetic men, and μ_f = mean androstenedione level for the population of diabetic women. We are interested in comparing the population means given the observed data, i.e. testing $H_0 : \mu_m = \mu_f$.

The raw data is read into R and made into one large data vector \mathbf{Y} below. The boxplots suggest that the distributions are fairly symmetric and normality within populations seems reasonable.

However, the assumption of equal population standard deviations is unreasonable. The sample standard deviation for men is noticeably larger than the women's standard deviation, even with outliers in the women's sample. A test of $H_0 : \sigma_m = \sigma_f$ is rejected at the 1% level.

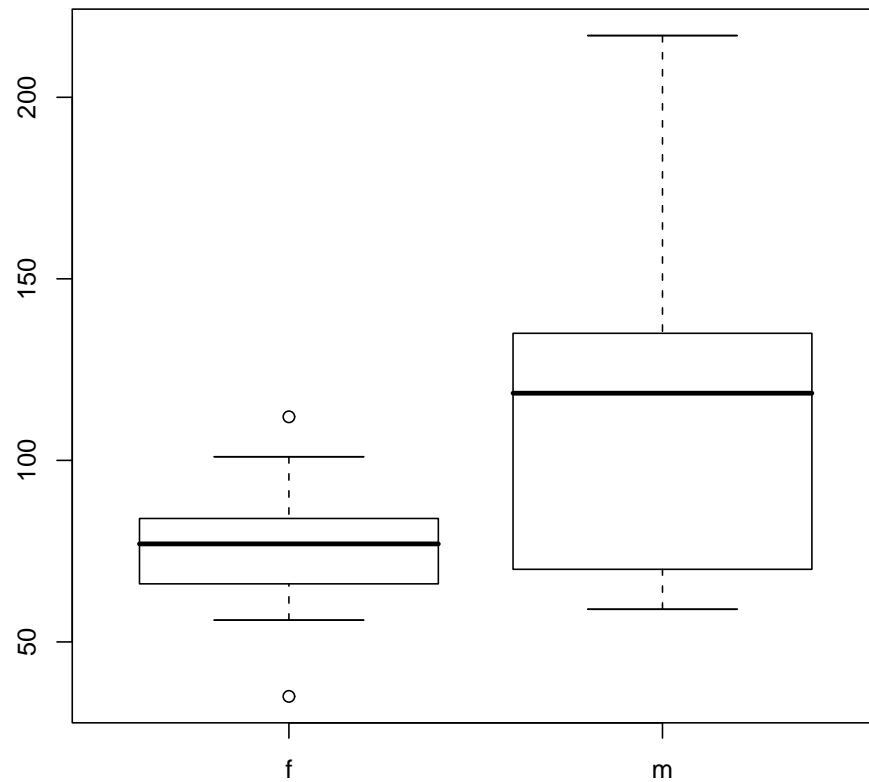


Figure 1: The assumption $\sigma_f = \sigma_m$ is probably not reasonable.

```

> f <- c(84,87,77,84,73,66,70,35,77,73,56,112,56,84,80,101,66,84)
> m <- c(217,123,80,140,115,135,59,126,70,63,147,122,108,70)
> y <- c(f,m)
> g <- factor(c(rep("f",length(f)),rep("m",length(m))))
> g
 [1] f f f f f f f f f f f f f f f f m m m m m m m m m m m m m m
> boxplot(y~g)
> levene.test(y,g)
      Classical Levene's test based on the absolute deviations from the mean
data:  y Test Statistic = 7.947, p-value = 0.00845

```

Although the normality assumption seems fine (to me), there are two outliers among the women. If we are worried about it, we can formally test for normality. Recall that accepting the null $H_0 : \text{data are normal}$ does not necessarily mean the data *really are* normal, but rather that we don't have evidence to the contrary. We do not find evidence (below) that the data are non-normal in either group.

```
> ad.test(f)
      Anderson-Darling normality test
data:  f A = 0.3947, p-value = 0.3364
> ad.test(m)
      Anderson-Darling normality test
data:  m A = 0.4718, p-value = 0.2058
```

Given that we are comfortable with the assumption that data are normal within groups, but that the standard deviations are different among men and women, the Satterthwaite analysis, also called the Welch test, is more appropriate here than the pooled variance analysis.

```
> t.test(y~g)
      Welch Two Sample t-test
data:  y by g t = -3.0235, df = 16.295, p-value = 0.007946
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -62.33778 -10.99555
sample estimates: mean in group f mean in group m
      75.83333      112.50000
```

With a p -value of 0.008 we reject $H_0 : \mu_f = \mu_m$ at the 1% level, strong evidence that the true mean androstenedione is different across gender.

There is a normality test built in to R that does not need to be loaded as does `ad.test()` (from the package `nortest`). The Shapiro-Wilk test for normality is a commonly used alternative test. The Anderson-Darling tests looks primarily for evidence of non-normal data in the tails of a distribution; the Shapiro-Wilk emphasizes lack of symmetry in the distribution; i.e. less emphasis placed on the tails.

```
> shapiro.test(f)
      Shapiro-Wilk normality test
data:  f W = 0.9598, p-value = 0.5969
> shapiro.test(m)
      Shapiro-Wilk normality test
data:  m W = 0.9059, p-value = 0.1376
```

The Shapiro-Wilk test gives quite different p -values but both tests accept the null hypotheses that the two data samples are normal.

If normality seems questionable, often a transformation of the data might yield more “normal looking” data. The $\log(\cdot)$ function is typically used for data that are skewed right. If a t -test is performed on the means of the log-transformed data, this results in a test for differences in medians for the original data.

If a transformation to approximate normality cannot be found, but the two distributions have roughly the same overall shape, the nonparametric Mann-Whitney test for differences in location can be used straightaway.

See example 11.2.1.1 (pp. 428–433, 442–443) for an in-depth analysis of a two-sample problem involving transformations.

Mann Whitney nonparametric test Section 11.2.3, pp. 435–442.

The Mann-Whitney test assumes

$$Y_{11}, \dots, Y_{1n_1} \stackrel{iid}{\sim} F_1 \text{ independent } Y_{21}, \dots, Y_{2n_2} \stackrel{iid}{\sim} F_2,$$

where F_1 is the cdf of data from the first group and F_2 is the cdf of data from the second group. The null hypothesis is $H_0 : F_1 = F_2$, i.e. that the distributions of data in the two groups are identical.

The alternative is $H_1 : F_1 \neq F_2$. One-sided tests can also be performed.

Although the test statistic is built assuming $F_1 = F_2$, the alternative is often taken to be that the population *medians* are unequal. This is a fine way to report the results of the test. Additionally, a CI will give a plausible range for Δ in the shift model $F_2(x) = F_1(x - \Delta)$. Δ can be the difference in medians or means.

The Mann-Whitney test is intuitive. The data are

$$y_{11}, y_{12}, \dots, y_{1n_1} \quad \text{and} \quad y_{21}, y_{22}, \dots, y_{2n_2}.$$

For each observation j in the first group count the number of observations in the second group c_j that are smaller; ties result in adding 0.5 to the count.

Assuming H_0 is true, on average half the observations in group 2 would be above Y_{1j} and half would be below if they come from the same distribution. That is $E(c_j) = 0.5n_2$.

The sum of these guys is $U = \sum_{j=1}^{n_1} c_j$ and has mean $E(U) = 0.5n_1n_2$. The variance is a bit harder to derive, but is $\text{Var}(U) = n_1n_2(n_1 + n_2 + 1)/12$.

Something akin to the CLT tells us

$$Z_0 = \frac{U - E(U)}{\sqrt{n_1 n_2 (n_1 + n_2 + 1) / 12}} \overset{\bullet}{\sim} N(0, 1),$$

when H_0 is true. Seeing a U far away from what we expect under the null gives evidence that H_0 is false; U is then standardized as usual (subtract off then mean we expect under the null and standardize by an estimate of the standard deviation of U). A p -values can be computed as usual as well as a CI.

Example: cooling rates of meteorites (courtesy E. Bedrick)

One theory of the formation of our solar system states that all solar system meteorites have the same evolutionary history and thus have the same cooling rates. By a delicate analysis based on measurements of phosphide crystal widths and phosphide-nickel content, the cooling rates, in degrees Celsius per million years, were determined for samples taken from meteorites named in the accompanying table after where they were found.

Walker County	0.69	0.23	0.10	0.03	0.56	0.10	0.01	0.02	0.04	0.22		
Uwet	0.21	0.25	0.16	0.23	0.47	1.20	0.29	1.10	0.16			
Tocopilla	5.60	2.70	6.20	2.90	1.50	4.00	4.30	3.00	3.60	2.40	6.70	3.80

We will explore whether the Walker County and Uwet meteorites have the same cooling rate and consider all three meteorites simultaneously a bit later on.

```

> uwet <- c(0.21,0.25,0.16,0.23,0.47,1.2,0.29,1.1,0.16)
> walk <- c(0.69,0.23,0.10,0.03,0.56,0.1,0.01,0.02,0.04,0.22)
> y <- c(uwet,walk)
> g <- factor(c(rep("u",length(uwet)),rep("w",length(walk))))
> boxplot(y~g)
> wilcox.test(y~g,conf.int=TRUE)
      Wilcoxon rank sum test with continuity correction
data:  y by g W = 69.5, p-value = 0.04974
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.0000449737 0.4499654518
sample estimates: difference in location
                0.1713625

```

Let η_1 and η_2 be the median cooling rates in the Uwet and Walker meteorites respectively. We estimate $\hat{\Delta} = \hat{\eta}_1 - \hat{\eta}_2 = 0.171$. We reject $H_0 : \eta_1 = \eta_2$ at the 5% level (just barely!) We are 95% confident that the median cooling rate for the Uwet meteorite is between 0.000045 and 0.4500 larger than for Walker.

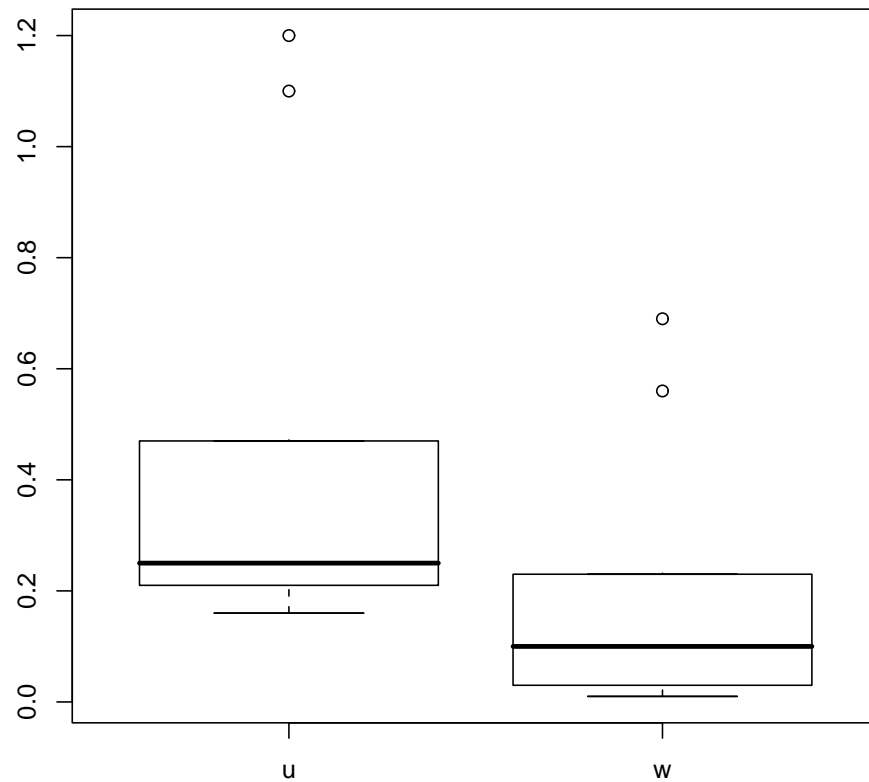


Figure 2: These are essentially shifted versions of each other.

We rejected H_0 , but it was a close call. Another option would be to try transforming the data to get approximately normal samples. The natural log is often used for data that are skewed toward larger values such as these data. Although we reject normality for the untransformed data, normality is not rejected for the log-transformed data. A t -test on the log-transformed data yields a p -value of 0.026. The normal-based procedure provides a bit more power to reject the null in this case.

We can interpret this null as $H_0 : \eta_1 = \eta_2$ as before. I will show this on the board.

Given that the nonparametric test makes no assumptions on the probability distributions, I would choose reporting the results of the Mann-Whitney test here.

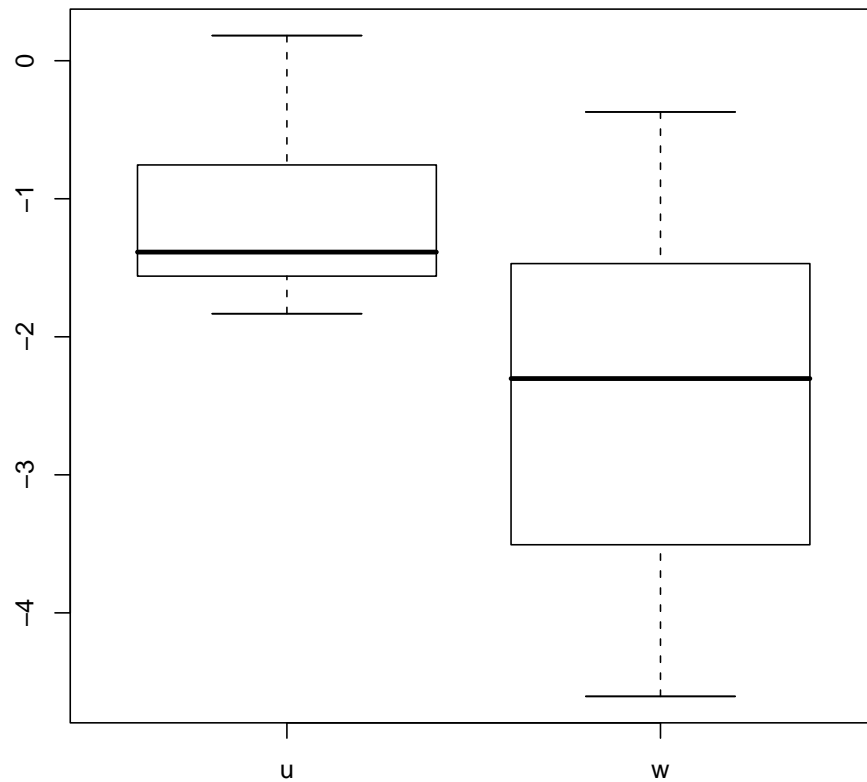


Figure 3: Log-transformed data not *too* “non-normal” looking.

```
> boxplot(log(y)~g)
> ad.test(uwet)
      Anderson-Darling normality test
data:  uwet A = 1.1898, p-value = 0.001987
> ad.test(walk)
      Anderson-Darling normality test
data:  walk A = 0.9622, p-value = 0.008967
> ad.test(log(uwet))
      Anderson-Darling normality test
data:  log(uwet) A = 0.6125, p-value = 0.07554
> ad.test(log(walk))
      Anderson-Darling normality test
data:  log(walk) A = 0.2004, p-value = 0.8354
> t.test(log(y)~g)
      Welch Two Sample t-test
data:  log(y) by g t = 2.4959, df = 14.121, p-value = 0.02555
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1823032 2.3970720
```

Comments:

- The t -test with equal variances provides greater power to reject H_0 when the variances truly are equal, but the power gain is minimal.
- The Satterthwaite approach is correct when the variances are unequal. When unsure, this is more appropriate.
- When data do not appear to be normal, the Mann-Whitney nonparametric test may be preferred.
- The Mann-Whitney test is robust to outlying observations and is invariant to monotone transformations of the data (e.g. taking the log of each observation does not change the p -value).