Statistics 2 Unit 5

Kurt Hornik

Outline

- Comparing two samples
- Analysis of categorical data
- Summary

Outline

- Comparing two samples
  - Comparing two independent samples
    - Methods based on the normal distribution
    - A nonparametric method: The Mann-Whitney test
  - Comparing paired samples
    - Methods based on the normal distribution
    - A nonparametric method: The signed rank test
- Analysis of categorical data
- Summary

Motivation

Suppose

- a sample $X_1, \ldots, X_n$ is drawn from a normal distribution with mean μ_X and variance σ^2
- an independent sample $Y_1, \ldots, Y_m$ is drawn from a normal distribution with mean μ_Y and variance σ^2.

We are interested in $\mu_X - \mu_Y$.

This is "naturally" estimated by $\bar{X} - \bar{Y}$.

Clearly,

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_X - \mu_Y$$

and by independence,

$$\mathrm{var}(\bar{X} - \bar{Y}) = \mathrm{var}(\bar{X}) + \mathrm{var}(-\bar{Y}) = \mathrm{var}(\bar{X}) + \mathrm{var}(\bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m}.$$

Motivation

Hence by normality of the samples,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

If σ^2 were known, a confidence interval for $\mu_X - \mu_Y$ could be based on

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

which has a standard normal distribution, giving

$$(\bar{X} - \bar{Y}) \pm z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Motivation

In general, σ^2 will not be known and must be estimated, e.g., by using the **pooled sample variance**

$$
\begin{aligned}
s_p^2 &= \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2} \\
&= \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{j=1}^{m}(Y_j-\bar{Y})^2\right).
\end{aligned}
$$

We then have the following result.

Two-sample t tests

Theorem. *Suppose that $X_1, \ldots, X_n$ are independent and normally distributed random variables with mean μ_X and variance σ^2, and that $Y_1, \ldots, Y_m$ are independent and normally distributed random variables with mean μ_Y and variance σ^2, and that the Y_j are independent of the X_i. Then the statistic*

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

follows a t distribution with $m + n - 2$ degrees of freedom. □

Two-sample t tests

Let

$$s_{\bar{X}-\bar{Y}} = s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

denote the estimated standard deviation of $\bar{X} - \bar{Y}$.

Two-sample t tests

Let

$$s_{\bar{X}-\bar{Y}} = s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

denote the estimated standard deviation of $\bar{X} - \bar{Y}$.

Corollary. *Under the above assumptions, a* $100(1-\alpha)\%$ *confidence interval for* $\mu_X - \mu_Y$ *is*

$$(\bar{X} - \bar{Y}) \pm t_{m+n-2,1-\alpha/2} s_{\bar{X}-\bar{Y}}.$$

□

Example: Ice

Two methods, A and B, were used to determine the latent heat of fusion of ice (Natrella, 1963).

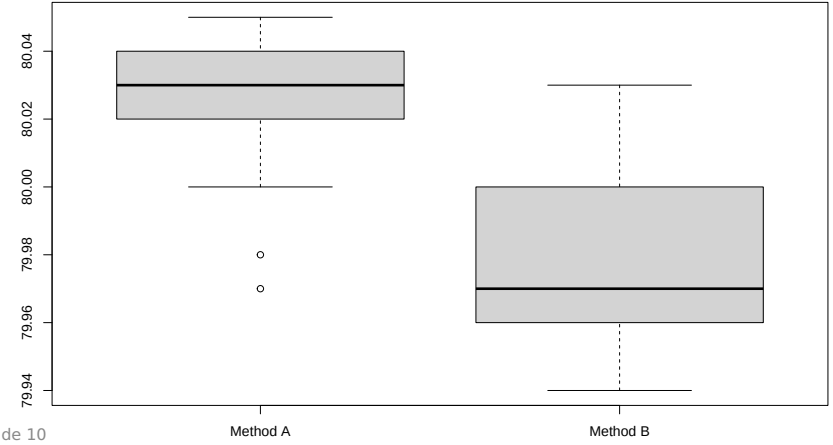
Measurements were obtained for the change in total heat from ice at $-72°$ C to water at $0°$ C in calories per gram of mass:

```
R> A <- scan("Data/icea.txt")
R> A

 [1] 79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02
[12] 80.00 80.02

R> B <- scan("Data/iceb.txt")
R> B

[1] 80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97
```

Example: Ice

It is fairly obvious that there is a difference between the methods:

```
R> boxplot(A, B, names = c("Method A", "Method B"))
```



Example: Ice

Doing computations by hand:

```
R> n_A <- length(A); m_A <- mean(A); s_A <- sd(A)
R> c(n_A, m_A, s_A)

[1] 13.00000000 80.02076923  0.02396579

R> n_B <- length(B); m_B <- mean(B); s_B <- sd(B)
R> c(n_B, m_B, s_B)

[1]  8.00000000 79.97875000  0.03136764

R> s_p <- sqrt(((n_A - 1) * s_A^2 + (n_B - 1) * s_B^2) /
+              (n_A + n_B - 2))
R> s_p

[1] 0.02693052
```

Example: Ice

This gives the following estimates for $\bar{X} - \bar{Y}$ and $s_{\bar{X}-\bar{Y}}$:

```
R> Delta <- m_A - m_B
R> Delta
```

```
[1] 0.04201923
```

```
R> s_Delta <- s_p * sqrt(1 / n_A + 1 / n_B)
R> s_Delta
```

```
[1] 0.01210146
```

Example: Ice

With the *t* quantile

```
R> q <- qt(0.975, n_A + n_B - 2)
```

we finally get the confidence interval

```
R> c(Delta - q * s_Delta, Delta + q * s_Delta)
[1] 0.01669058 0.06734788
```

Example: Ice

But because we have R, we can simply use function `t.test` to obtain the confidence interval and corresponding hypothesis test:

```
R> t.test(A, B, var.equal = TRUE)

        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Two-sample t tests

Rice (page 426f) shows that this t test is actually the LRT in the case of unknown but equal variances in the two samples.

(So the “natural” idea can also be obtained from first principles.)

Two-sample t tests

Rice (page 426f) shows that this t test is actually the LRT in the case of unknown but equal variances in the two samples.

(So the “natural” idea can also be obtained from first principles.)

In fact, for hypothesis testing for the two-sample problem, there are three common alternative hypotheses to $H_0 : \mu_X = \mu_Y$:

$$H_1 : \mu_X \neq \mu_Y, \qquad H_2 : \mu_X < \mu_Y, \qquad H_3 : \mu_X > \mu_Y.$$

The first is a **two-sided alternative**, the other two are **one-sided alternatives**.

Two-sample t tests

Hypothesis tests for all three alternatives are based on the t statistic

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}$$

with rejection regions

$$H_1 : |t| > t_{m+n-2,1-\alpha/2}, \qquad H_2 : t < t_{m+n-2,\alpha}, \qquad H_3 : t > t_{m+n-2,1-\alpha}$$

(and corresponding two- or one sided confidence intervals).

Example: Ice

E.g., to test against the alternative $\mu_A > \mu_B$ for the ice data,

```
R> t.test(A, B, alternative = "greater", var.equal = TRUE)

        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.001276
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0210942       Inf
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Two-sample t tests

If the variances are not known to be equal, a natural estimate of $\text{var}(\bar{X} - \bar{Y})$ is

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

If this is used in the denominator of the test statistic, the t distribution no longer holds exactly, but approximately with

$$\frac{(s_X^2/n + s_Y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_Y^2/m)^2/(m-1)}$$

degrees of freedom: so-called **Welch** approximation to the sampling distribution.

Two-sample t tests

This is actually the default in R:

```
R> t.test(A, B)

        Welch Two Sample t-test

data:  A and B
t = 3.2499, df = 12.027, p-value = 0.006939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01385526 0.07018320
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Motivation

Nonparametric methods do not assume the data follow a particular distributional form.

(So we're moving outside the framework of traditional parametric inference we considered thus far.)

Often, data are replaced by ranks, making results invariant under monotonic transformations, and moderating the influence of outliers.

Suppose we have two independent sample $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from probability distributions F and G, respectively, and that it is desired to test the null hypothesis that $F = G$.

Motivation

We will develop the Mann-Whitney test, also known as the **Wilcoxon rank sum test**.

This is based on the idea that under the null, assigning the pooled (sorted) observations to the samples is “random” in the sense that all assignments are equiprobable.

Consider a simple example. Suppose observations are

$$X : 1, 3 \qquad Y : 4, 6.$$

As $1 < 3 < 4 < 6$, the corresponding ranks (in the pooled data) are

$$X : 1, 2 \qquad Y : 3, 4$$

with rank sums 3 and 7, respectively.

Motivation

Now, under the null, every assignment of the ranks to the samples is equally likely. For the ranks of the second group, we have

Ranks	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
R	3	4	5	5	6	7

and thus

$$\mathbb{P}(R = 7) = 1/6.$$

Motivation

Now, under the null, every assignment of the ranks to the samples is equally likely. For the ranks of the second group, we have

Ranks	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
R	3	4	5	5	6	7

and thus

$$\mathbb{P}(R = 7) = 1/6.$$

In the general case, under the null every possible assignment of the $m + n$ ranks to the n elements of the second group is equally likely.

Example: Ice

We compute the ranks for methods *A* and *B*:

```
R> C <- c(A, B)
R> r_A <- rank(C)[seq_along(A)]
R> r_B <- rank(C)[seq_along(B) + length(A)]
```

Note how ties are handled.

The rank sum of the smaller sample is

```
R> sum(r_B)
```

```
[1] 51
```

The Mann-Whitney Test

What about the distribution of the rank sums under the null? We can use R:

```
R> wilcox.test(A, B)

        Wilcoxon rank sum test with continuity correction

data:  A and B
W = 89, p-value = 0.007497
alternative hypothesis: true location shift is not equal to 0
```

We note 2 things:

- a warning about exact *p*-values and ties (not shown here).

The Mann-Whitney Test

What about the distribution of the rank sums under the null? We can use R:

```
R> wilcox.test(A, B)

        Wilcoxon rank sum test with continuity correction

data:  A and B
W = 89, p-value = 0.007497
alternative hypothesis: true location shift is not equal to 0
```

We note 2 things:

- a warning about exact *p*-values and ties (not shown here).
- the value of the test statistic, which is not the sum of the ranks in the smaller sample as in Rice. What R uses, is the symmetric version of the test statistic (see below)

The Mann-Whitney Test

Let T_Y denote the sum of the ranks of $Y_1, \ldots, Y_m$.

Theorem. *If $F = G$,*

$$\mathbb{E}(T_Y) = \frac{m(m+n+1)}{2}, \qquad \mathrm{var}(T_Y) = \frac{mn(m+n+1)}{12}.$$

□

What R actually does is compute the rank sum for the **first** sample which would have expectation $n(m+n+1)/2$ and subtract $n(n+1)/2$:

```
R> sum(r_A) - n_A * (n_A + 1) / 2
```

```
[1] 89
```

which has expectation $mn/2$ which is symmetric in m and n.

The Mann-Whitney Test

If the samples are interchanged, R would use

```
R> sum(r_B) - n_B * (n_B + 1) / 2
```

```
[1] 15
```

as can be verified by inspection:

```
R> wilcox.test(B, A)$statistic
```

```
 W
15
```

The Mann-Whitney Test

The Mann-Whitney (Wilcoxon rank sum) test can also be derived as follows.

Consider estimating

$$\pi = \mathbb{P}(X < Y).$$

The Mann-Whitney Test

The Mann-Whitney (Wilcoxon rank sum) test can also be derived as follows.

Consider estimating

$$\pi = \mathbb{P}(X < Y).$$

(This for simplicity assumes continuous distributions. In general, one needs

$$\pi = \mathbb{P}(X < Y) + \mathbb{P}(X = Y)/2$$

with the indicators below modified accordingly.)

The Mann-Whitney Test

The Mann-Whitney (Wilcoxon rank sum) test can also be derived as follows.

Consider estimating

$$\pi = \mathbb{P}(X < Y).$$

(This for simplicity assumes continuous distributions. In general, one needs

$$\pi = \mathbb{P}(X < Y) + \mathbb{P}(X = Y)/2$$

with the indicators below modified accordingly.)

A natural estimate would be

$$\hat{\pi} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}, \qquad Z_{ij} = 1_{X_i < Y_j}.$$

The Mann-Whitney Test

Now note that

$$\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij}, \qquad V_{ij} = 1_{X_{(i)}<Y_{(j)}}.$$

If the rank of $Y_{(j)}$ in the pooled sample is denoted by $R_{Y,j}$, then the number of X less than $Y_{(j)}$ is $R_{Y,j} - j$ (in the case of no ties), hence

$$\sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij} = \sum_{j=1}^{m}(R_{Y,j} - j) = T_Y - m(m+1)/2 = U_Y.$$

giving the symmetric version of the test statistic.

The Mann-Whitney Test

Using this notation, R's W is really U_X. I.e., $\hat{\pi}$ is

```
R> wilcox.test(B, A)$statistic / (n_A * n_B)

        W
0.1442308
```

The Mann-Whitney Test

Note that this is not the same as

```
R> mean(outer(A, B, `<`))
```

```
[1] 0.09615385
```

as the example actually has ties:

```
R> c(sum(outer(A, B, `<`)), sum(outer(A, B, `<=`)))
```

```
[1] 10 20
```

Argh. Mutatis mutandis . . .

The Mann-Whitney Test

Corollary. *If $F = G$,*

$$\mathbb{E}(U_Y) = \frac{mn}{2}, \qquad \text{var}(U_Y) = \frac{mn(m+n+1)}{2}.$$

□

The Mann-Whitney Test

The Mann-Whitney test can be inverted to obtain confidence intervals for location shifts: consider the shift model $G(x) = F(x - \Delta)$.

Then for testing the null that the shift parameter is Δ, we can use

$$U_Y(\Delta) = \#\{(i,j) : X_i - (Y_j - \Delta) < 0\} = \#\{(i,j) : Y_j - X_i > \Delta\}.$$

One can show that the distribution of $U_Y(\Delta)$ is symmetric about $mn/2$.

A $100(1-\alpha)\%$ confidence interval for Δ is thus of the form

$$C = \{\Delta : k \leq U_Y(\Delta) \leq mn - k\}$$

which can be rewritten in terms of the ordered $X_i - Y_j$.

The Mann-Whitney Test

In R, confidence intervals are obtained via `conf.int = TRUE`:

```
R> wilcox.test(A, B, conf.int = TRUE)

        Wilcoxon rank sum test with continuity correction

data:  A and B
W = 89, p-value = 0.007497
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.01000082 0.07001754
sample estimates:
difference in location
            0.05008264
```

The Mann-Whitney Test

Bootstrap for the two-sample problem: suppose again that $\pi = \mathbb{P}(X < Y)$ is estimated by $\hat{\pi}$.

How can the standard error of this be estimated?

(Note that the confidence intervals are computed under the assumption that $F = G$.)

If F and G were known, we could generate bootstrap samples and compute $\hat{\pi}_1, \ldots, \hat{\pi}_B$ from these.

As they are not known, one instead uses the empirical distributions F_n and G_m.

I.e., one repeatedly randomly selects n values from the observed $X_1, \ldots, X_n$ with replacement, and m values from the observed $Y_1, \ldots, Y_m$, and calculates the resulting $\hat{\pi}$, generating a bootstrap sample $\hat{\pi}_1, \ldots, \hat{\pi}_B$.

The Mann-Whitney Test

This is our first example of a **non-parametric bootstrap**, which is based on suitably resampling the observations.

Outline

- Comparing two samples
  - Comparing two independent samples
    - Methods based on the normal distribution
    - A nonparametric method: The Mann-Whitney test
  - Comparing paired samples
    - Methods based on the normal distribution
    - A nonparametric method: The signed rank test
- Analysis of categorical data
- Summary

Motivation

Often, samples are paired, e.g., by matching cases to controls and then randomly assigning to treatment and control groups, or by taking “before” and “after” measurements on the same object.

Given pairing, the samples are no longer independent.

Motivation

Often, samples are paired, e.g., by matching cases to controls and then randomly assigning to treatment and control groups, or by taking "before" and "after" measurements on the same object.

Given pairing, the samples are no longer independent.

Denote pairs as (X_i, Y_i) where the X and Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively.

Assume that different pairs are independent with $\text{cov}(X_i, Y_i) = \sigma_{XY} = \rho\sigma_X\sigma_Y$, where ρ is the correlation of X and Y.

Then the differences $D_i = X_i - Y_i$ are independent with

$$\mathbb{E}(D_i) = \mu_X - \mu_Y, \qquad \text{var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y.$$

Motivation

Suppose the parameter of interest is

$$\mu_X - \mu_Y.$$

For the natural estimate $\bar{D} = \bar{X} - \bar{Y}$,

$$\mathbb{E}(\bar{D}) = \mu_X - \mu_Y, \qquad \text{var}(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y).$$

Motivation

Compare to the independent case: if $\rho > 0$, the variance of $\bar{D}$ is smaller.

In general, if $\sigma_X^2 = \sigma_Y^2 = \sigma^2$,

$$\text{var}(\bar{D}) = \frac{1}{n}(2\sigma^2 - 2\rho\sigma^2) = \frac{2\sigma^2}{n}(1-\rho).$$

If the X and Y were independent (as previously in the 2-sample case),

$$\text{var}(\bar{X} - \bar{Y}) = \text{var}(\bar{X}) + \text{var}(\bar{Y}) = \frac{2\sigma^2}{n}$$

(corresponding to $\rho = 0$, of course).

Paired t test

If the differences have a normal distribution with

$$\mathbb{E}(D_i) = \mu_D = \mu_X - \mu_Y, \qquad \text{var}(D_i) = \sigma_D^2,$$

with σ_D^2 typically unknown, inference will be based on

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

which follows a t distribution with $n - 1$ degrees of freedom.

Paired t test

If the differences have a normal distribution with

$$\mathbb{E}(D_i) = \mu_D = \mu_X - \mu_Y, \qquad \text{var}(D_i) = \sigma_D^2,$$

with σ_D^2 typically unknown, inference will be based on

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

which follows a t distribution with $n-1$ degrees of freedom.

In particular, if $H_0 : \mu_D = 0$, then under H_0,

$$t = \frac{\bar{D}}{s_{\bar{D}}} \sim t_{n-1}.$$

Example: Smoking

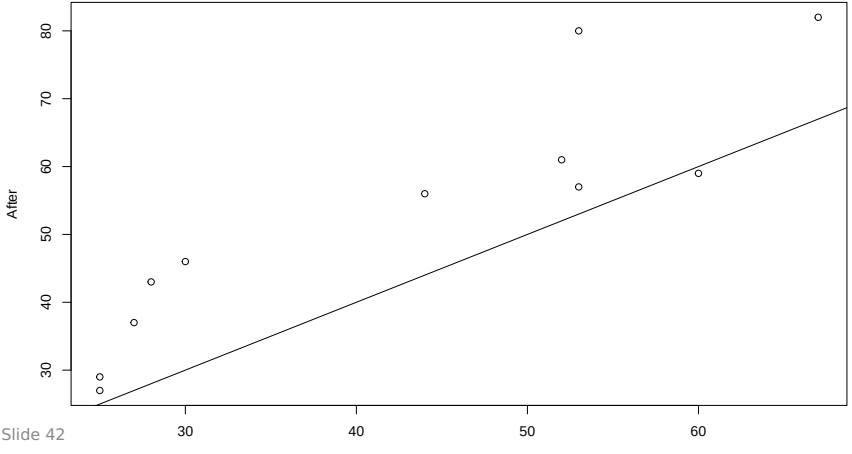
Levine (1973) drew blood samples from 11 individuals before and after smoking and measured the extent to which the blood platelets aggregated.

```
R> platelet <- read.table("Data/platelet.txt", sep = ",",
+                         header = TRUE)
R> B <- platelet$before
R> A <- platelet$after
```

Example: Smoking

We can inspect the data via

```
R> plot(B, A, xlab = "Before", ylab = "After"); abline(0, 1)
```



Example: Smoking

Note that

```
R> cor(B, A)
[1] 0.9012976
```

so pairing is quite efficient.

Example: Smoking

Inference can be performed “by hand”, using

```
R> D <- B - A
R> t <- mean(D) / (sd(D) / sqrt(length(D)))
R> t

[1] -4.271609
```

Example: Smoking

In R, we can use `t.test()` with argument `paired = TRUE`:

```
R> t.test(B, A, paired = TRUE)

        Paired t-test

data:  B and A
t = -4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -15.63114  -4.91431
sample estimates:
mean difference
      -10.27273
```

The signed rank test

The signed rank test (also known as the **Wilcoxon signed rank test**) is based on a simple idea.

Compute the differences $D_i = X_i - Y_i$, rank the absolute values of the D_i, and compute the sum of the ranks for which the differences are positive.

In our example,

```
R> D <- B - A
R> R <- rank(abs(D))
R> sum(R[D > 0])

[1] 1
```

The signed rank test

Intuitively, if there was no difference between the paired variables, about half of the D_i should be positive, and the signed rank sum should not be too extreme (small or large).

More precisely, consider the null hypothesis that the distribution of D is symmetric about 0. If this distribution is continuous, then under H_0, all sign combinations have equal probability $1/2^n$.

The signed rank sum is then of the form

$$W_+ = \sum_{k=1}^{n} kI_k,$$

where I_k is the indicator that the k-th largest $|D_i|$ has positive sign.

The signed rank test

Under H_0, the I_k are i.i.d. Bernoulli with $p = 1/2$, so $\mathbb{E}(I_k) = 1/2$, $\mathrm{var}(I_k) = 1/4$,

$$\mathbb{E}(W_+) = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4}, \qquad \mathrm{var}(W_+) = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24}.$$

In particular,

$$\mathbb{P}(W_+ = 1) = \mathbb{P}(I_1 = 1, I_2 = \cdots = I_n = 0) = 1/2^n.$$

The signed rank test

In our example

```
R> 1 / 2^(length(D))
[1] 0.0004882812
```

which is the same as

```
R> dsignrank(1, length(D))
[1] 0.0004882812
```

and rejecting if W_+ is extremely small or large would have p-value

```
R> 2 * psignrank(1, length(D))
[1] 0.001953125
```

The signed rank test

In case of ties (as in our case), things are a bit messier, and e.g. the normal approximation based on the above mean and variance is used. Compactly:

```
R> wilcox.test(B, A, paired = TRUE)

        Wilcoxon signed rank test with continuity correction

data:  B and A
V = 1, p-value = 0.005056
alternative hypothesis: true location shift is not equal to 0
```

The signed rank test

One can also invert this test to obtain confidence intervals for the **pseudomedian**.

The pseudomedian of a probability distribution F is the median of the distribution of $(U + V)/2$, where U and V are independent with distribution F.

If F is symmetric, then median and pseudomedian coincide.

The signed rank test

In our example, this gives

```
R> wilcox.test(B, A, paired = TRUE, conf.int = TRUE)

        Wilcoxon signed rank test with continuity correction

data:  B and A
V = 1, p-value = 0.005056
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -15.499990  -4.000002
sample estimates:
(pseudo)median
     -9.500014
```

Outline

- Comparing two samples
- Analysis of categorical data
- Summary

Outline

- Comparing two samples
  - Methods based on the normal distribution
  - A nonparametric method: The Mann-Whitney test
  - Methods based on the normal distribution
  - A nonparametric method: The signed rank test
- Analysis of categorical data
  - Fisher's exact test
  - The chi-squared test of homogeneity
  - The chi-squared test of independence
  - Matched-pairs designs

Motivation

Rosen and Jordan (1974) experiment with male bank supervisors attending a management institute. In one experiment, supervisors were given a personnel file and had to decide whether to promote the employee or to hold the file and interview additional candidates. By random selection, 24 supervisors examined files labeled as from a male and 24 files labeled as from a female employee; files were otherwise identical.

Motivation

Rosen and Jordan (1974) experiment with male bank supervisors attending a management institute. In one experiment, supervisors were given a personnel file and had to decide whether to promote the employee or to hold the file and interview additional candidates. By random selection, 24 supervisors examined files labeled as from a male and 24 files labeled as from a female employee; files were otherwise identical.

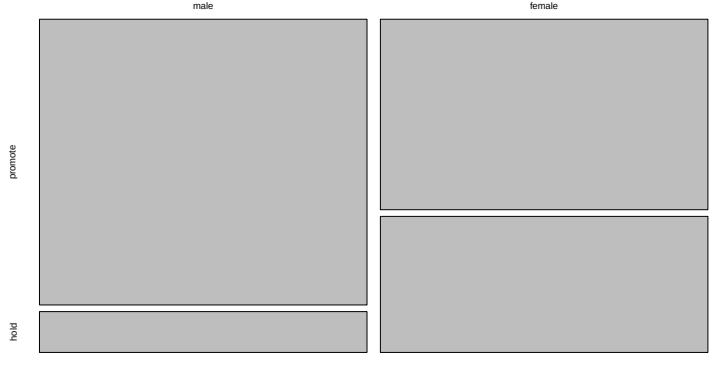
Results were as follows.

	Male	*Female*
Promote	*21*	*14*
Hold File	*3*	*10*

Is there evidence for a gender bias?

Motivation

A visual comparison of the frequencies:



Fisher's exact test

Under the null of no bias, the observed "differences" would be due only to the random assignment of supervisors to files. We denote the counts in the table and the margins as follows:

N_{11}	N_{12}	$n_{1\cdot}$
N_{21}	N_{22}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

The dots are hard to see/read at first encouter: they simply indicate summing out (taking margins).

Fisher's exact test

Under the null of no bias, the observed "differences" would be due only to the random assignment of supervisors to files. We denote the counts in the table and the margins as follows:

N_{11}	N_{12}	$n_{1\cdot}$
N_{21}	N_{22}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

The dots are hard to see/read at first encouter: they simply indicate summing out (taking margins).

According to the null hypothesis, the margins are **fixed**: the process of randomization determines the random fluctuation of the cell counts in the interior of the table subject to the constraints of the margin.

Fisher's exact test

With these constraints, there is in fact only 1 degree of freedom in the interior.

Consider the count N_{11}.

Under H_0, this is distributed as the number of successes in 24 draws without replacement from a population of 35 successes and 13 failures, i.e., it has a hypergeometric distribution

$$\mathbb{P}(N_{11} = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}.$$

Fisher's exact test

In our case, the number of successes must be between 11 and 24:

```
R> round(dhyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0036 0.0206 0.0720 0.1620 0.2415 0.2415 0.1620
[10] 0.0720 0.0206 0.0036 0.0003 0.0000
```

Fisher's exact test

In our case, the number of successes must be between 11 and 24:

```
R> round(dhyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0036 0.0206 0.0720 0.1620 0.2415 0.2415 0.1620
[10] 0.0720 0.0206 0.0036 0.0003 0.0000
```

The null would be rejected for small or large values of n_{11}, e.g., for significance level $\alpha = 0.05$:

```
R> round(phyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0039 0.0245 0.0965 0.2585 0.5000 0.7415 0.9035
[10] 0.9755 0.9961 0.9997 1.0000 1.0000
```

suggests rejecting when $n_{11} \leq 14$ or $n_{11} \geq 21$.

Fisher's exact test

In our case, the number of successes must be between 11 and 24:

```
R> round(dhyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0036 0.0206 0.0720 0.1620 0.2415 0.2415 0.1620
[10] 0.0720 0.0206 0.0036 0.0003 0.0000
```

The null would be rejected for small or large values of n_{11}, e.g., for significance level $\alpha = 0.05$:

```
R> round(phyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0039 0.0245 0.0965 0.2585 0.5000 0.7415 0.9035
[10] 0.9755 0.9961 0.9997 1.0000 1.0000
```

suggests rejecting when $n_{11} \leq 14$ or $n_{11} \geq 21$.

In our case, $n_{11} = 21$ so the null of no bias is rejected at the 5% level.

Fisher's exact test

This is **Fisher's exact test**. In R,

```
R> tab <- matrix(c(21, 14, 3, 10), nrow = 2, byrow = TRUE)
R> fisher.test(tab)

        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.04899
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.00557 32.20580
sample estimates:
odds ratio
   4.83119
```

Fisher's exact test

Note that this is formulated in terms of the odds ratio corresponding to the table (but that the sample estimate is not the sample odds ratio).

R also provides suitable generalizations of Fisher's exact test to general $r \times c$ contingency tables.

Alternatively, function `r2dtable()` can be used for efficient generation of tables with given row and column margins, and hence for bootstrap versions of the test for independence of rows and columns given the margins.

Outline

- Comparing two samples
  - Methods based on the normal distribution
  - A nonparametric method: The Mann-Whitney test
  - Methods based on the normal distribution
  - A nonparametric method: The signed rank test
- Analysis of categorical data
  - Fisher's exact test
  - The chi-squared test of homogeneity
  - The chi-squared test of independence
  - Matched-pairs designs

Motivation

Suppose we have independent observations from J multinomial distributions with I cells each, and want to test the null that the cell probabilities of the multinomials are equal—i.e., test the homogeneity of the multinomial distributions.

Consider the following example from stylometry given in Morton (1978).

When Jane Austen died, she left the novel *Sanditon* only partially completed. A highly literate admirer finished the work based on the summary of the remainder, attempting to emulate Austen's style.

Motivation

The following table gives word counts obtained by Morton for Chapters from *Sense and Sensibility*, *Emma*, and *Sanditon* written by Austen (Sanditon I) and her admirer (Sanditon II):

Word	*Sense and Sensibility*	*Emma*	*Sanditon I*	*Sanditon II*
a	*147*	*186*	*101*	*83*
an	*25*	*26*	*11*	*29*
this	*32*	*39*	*15*	*15*
that	*94*	*105*	*37*	*22*
with	*59*	*74*	*28*	*43*
without	*18*	*10*	*10*	*4*
Total	*375*	*440*	*202*	*196*

Motivation

```
R> tab <-
+     matrix(c(147, 186, 101,  83,
+                25,  26,  11,  29,
+                32,  39,  15,  15,
+                94, 105,  37,  22,
+                59,  74,  28,  43,
+                18,  10,  10,   4),
+            ncol = 4, byrow = TRUE)
R> rownames(tab) <- c("a", "an", "this", "that", "with", "without")
R> colnames(tab) <- c("S&S", "Emma", "SandI", "SandII")
```

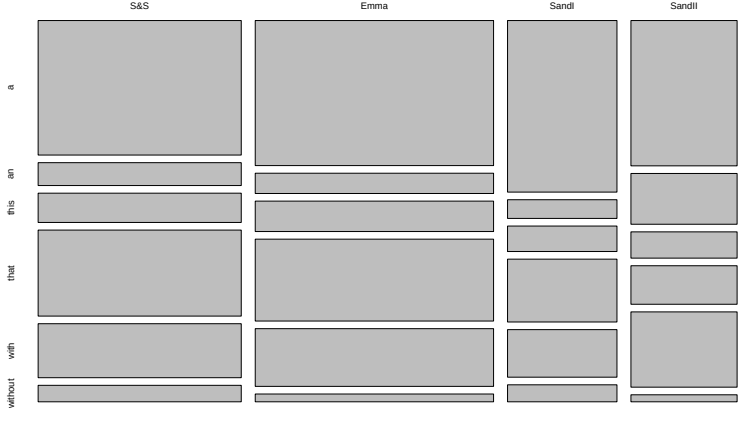
Motivation

```
R> tab

         S&S Emma SandI SandII
a        147  186   101     83
an        25   26    11     29
this      32   39    15     15
that      94  105    37     22
with      59   74    28     43
without   18   10    10      4
```

Motivation

A visual comparison of the frequencies:

```
R> mosaicplot(t(tab), main = "")
```



Motivation

We will use the following stochastic model:

- the counts for *Sense and Sensibility* will be modeled as a multinomial random variable with unknown cell probabilities and total count 375.
- the counts for `Emma` as multinomial with unknown cell probabilities and total count 440.
- etc.

Motivation

We will use the following stochastic model:

- the counts for *Sense and Sensibility* will be modeled as a multinomial random variable with unknown cell probabilities and total count 375.
- the counts for Emma as multinomial with unknown cell probabilities and total count 440.
- etc.

Write π_{ij} for the probability of category i in multinomial j. Then the null hypothesis is

$$H_0 : \pi_{i1} = \cdots = \pi_{ij}, \qquad i = 1, \ldots, I.$$

Write n_{ij} for the observed cell counts.

The chi-squared test of homogeneity

Theorem. *Under H_0 and independent multinomial sampling, the MLEs of the common cell probabilities π_i are*

$$\hat{\pi}_i = n_{i\cdot}/n_{\cdot\cdot}$$

The chi-squared test of homogeneity

Theorem. *Under H_0 and independent multinomial sampling, the MLEs of the common cell probabilities π_i are*

$$\hat{\pi}_i = n_{i\cdot}/n_{\cdot\cdot}$$

Proof. By independence,

$$\text{lik}(\pi_1, \ldots, \pi_I) = \prod_{j=1}^{J} \frac{n_{\cdot j}!}{n_{1j}! \cdots n_{Ij}!} \pi_1^{n_{1j}} \cdots \pi_I^{n_{Ij}} = \pi_1^{n_{1\cdot}} \cdots \pi_I^{n_{I\cdot}} \prod_{j=1}^{J} \frac{n_{\cdot j}!}{n_{1j}! \cdots n_{Ij}!}$$

The chi-squared test of homogeneity

To maximize, we use the Lagrangian

$$L(\pi_1, \ldots, \pi_I, \lambda) = \sum_{i=1}^{I} n_{i\cdot} \log(\pi_i) + \lambda \left(\sum_{i=1}^{I} \pi_i - 1 \right).$$

This has partials

$$\frac{\partial L}{\partial \pi_i} = \frac{n_{i\cdot}}{\pi_i} + \lambda.$$

Setting these to zero gives $\pi_i = -n_{i\cdot}/\lambda$ and thus the assertion by using the constraint. □

(We already did this.)

The chi-squared test of homogeneity

For multinomial j, the expected counts (under the null) are

$$E_{ij} = \hat{\pi}_{i} n_{\cdot j} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}.$$

Pearson's chi-squared statistic is therefore

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot} n_{\cdot j}/n_{\cdot\cdot}}$$

The chi-squared test of homogeneity

For large sample size, this approximately has a χ^2 distribution with

$$J(I-1)-(I-1)=(I-1)(J-1)$$

degrees of freedom, as

- each of the J multinomials has $I-1$ parameters
- under the null $I-1$ parameters are estimated.

The chi-squared test of homogeneity

For large sample size, this approximately has a χ^2 distribution with

$$J(I-1)-(I-1)=(I-1)(J-1)$$

degrees of freedom, as

- each of the J multinomials has $I-1$ parameters
- under the null $I-1$ parameters are estimated.

This is the **chi-squared test of homogeneity**.

Example: Austen

Using R, we first compare the frequencies for Austen's writings.

```
R> chisq.test(tab[, 1 : 3])

        Pearson's Chi-squared test

data:  tab[, 1:3]
X-squared = 12.271, df = 10, p-value = 0.2673
```

We can thus aggregate the Austen counts to one Austen "meta-novel":

```
R> tab <- cbind(Aus = rowSums(tab[, 1:3]), Imi = tab[, 4])
R> t(tab)

      a an this that with without
Aus 434 62   86  236  161      38
Imi  83 29   15   22   43       4
```

Example: Austen

A visual comparison of the frequencies:

```
R> mosaicplot(t(tab), main = "")
```

Example: Austen

We now test whether these counts follow the same frequency distribution:

```
R> cst <- chisq.test(tab)
R> cst

        Pearson's Chi-squared test

data:  tab
X-squared = 32.81, df = 5, p-value = 4.106e-06
```

Example: Austen

We now test whether these counts follow the same frequency distribution:

```
R> cst <- chisq.test(tab)
R> cst

        Pearson's Chi-squared test

data:  tab
X-squared = 32.81, df = 5, p-value = 4.106e-06
```

So the imitator was significantly unsuccessful!

Example: Austen

To see why, one can look at the contributions to the test statistic (the squared so-called Pearson residuals).

We can get these residuals with

```
R> res <- cst$residuals
```

Note that

```
R> sum(res^2)
```

```
[1] 32.80959
```

gives the X^2 test statistic.

Example: Austen

Looking at the residuals

```
R> round(res, 2)
```

```
           Aus   Imi
a         0.03 -0.06
an       -1.64  3.73
this      0.14 -0.33
that      1.34 -3.05
with     -0.77  1.75
without   0.47 -1.07
```

we see that Austen used *an* much less and *that* much more frequently than her imitator.

Outline

- Comparing two samples
  - Methods based on the normal distribution
  - A nonparametric method: The Mann-Whitney test
  - Methods based on the normal distribution
  - A nonparametric method: The signed rank test
- Analysis of categorical data
  - Fisher's exact test
  - The chi-squared test of homogeneity
  - The chi-squared test of independence
  - Matched-pairs designs

Motivation

In a demographic study of women listed in *Who's Who*, Kiser and Schaefer (1949) compiled the following table for 1436 women who were married at least once:

	Married once	*Married more*	*Total*
College	*550*	*61*	*611*
No College	*681*	*144*	*825*
Total	*1231*	*205*	*1436*

Is there a relationship between marital status and level of education?

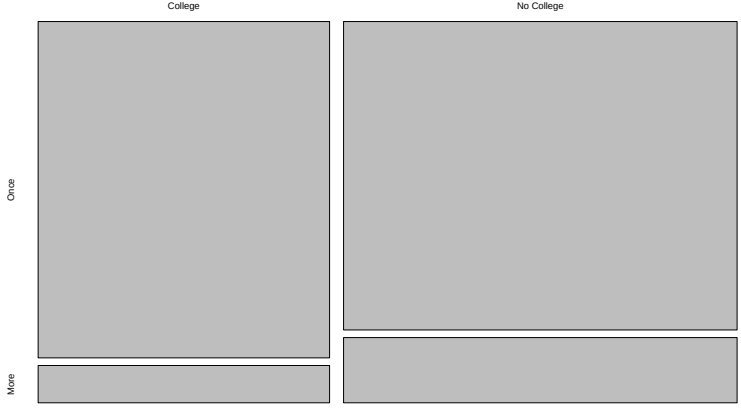
Motivation

```
R> tab <- matrix(c(550, 61, 681, 144), nrow = 2, byrow = TRUE)
R> rownames(tab) <- c("College", "No College")
R> colnames(tab) <- c("Once", "More")
R> tab

           Once More
College     550   61
No College  681  144
```

Motivation

A mosaic plot:

```
R> mosaicplot(tab, main = "")
```



The chi-squared test of independence

We model the data as coming from a sample of size n cross-classified in a table with I rows and J columns, a **contingency table**, with the joint distribution of the cell counts n_{ij} a multinomial with cell probabilities π_{ij}.

Note the difference to the previous section!

The chi-squared test of independence

We model the data as coming from a sample of size n cross-classified in a table with I rows and J columns, a **contingency table**, with the joint distribution of the cell counts n_{ij} a multinomial with cell probabilities π_{ij}.

Note the difference to the previous section!

If the row and column classifications are independent,

$$\pi_{ij} = \pi_{i.}\pi_{.j}.$$

The chi-squared test of independence

We model the data as coming from a sample of size n cross-classified in a table with I rows and J columns, a **contingency table**, with the joint distribution of the cell counts n_{ij} a multinomial with cell probabilities π_{ij}.

Note the difference to the previous section!

If the row and column classifications are independent,

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}.$$

We thus consider testing

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

versus the alternatives that the π_{ij} are free (apart from being non-negative with sum one).

The chi-squared test of independence

For cell probability matrix $[\pi_{ij}]$, the likelihood is

$$\text{lik}([\pi_{ij}]) = n! \prod_{i,j} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}.$$

The chi-squared test of independence

For cell probability matrix $[\pi_{ij}]$, the likelihood is

$$\text{lik}([\pi_{ij}]) = n! \prod_{i,j} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}.$$

Under H_0, the log-likelihood is thus

$$\ell = \log(n!) + \sum_{i,j} \left(n_{ij} \log(\pi_{i\cdot}\pi_{\cdot j}) - \log(n_{ij}!) \right)$$

The chi-squared test of independence

To find the MLEs under H_0, we can use the Lagrangian

$$L = \sum_{i,j} n_{ij} \big(\log(\pi_{i\cdot}) + \log(\pi_{\cdot j})\big) + \lambda \left(\sum_i \pi_{i\cdot} - 1 \right) + \mu \left(\sum_j \pi_{\cdot j} - 1 \right).$$

The chi-squared test of independence

To find the MLEs under H_0, we can use the Lagrangian

$$L = \sum_{i,j} n_{ij} \big(\log(\pi_{i\cdot}) + \log(\pi_{\cdot j})\big) + \lambda \left(\sum_i \pi_{i\cdot} - 1 \right) + \mu \left(\sum_j \pi_{\cdot j} - 1 \right).$$

This has partials

$$\frac{\partial L}{\partial \pi_{i\cdot}} = \sum_j \frac{n_{ij}}{\pi_{i\cdot}} + \lambda = \frac{n_{i\cdot}}{\pi_{i\cdot}} + \lambda, \qquad \frac{\partial L}{\partial \pi_{\cdot j}} = \sum_i \frac{n_{ij}}{\pi_{\cdot j}} + \mu = \frac{n_{\cdot j}}{\pi_{\cdot j}} + \mu.$$

By the usual computations, setting these to zero gives

$$\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n}, \qquad \hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

The chi-squared test of independence

Under H_0, the MLEs for the π_{ij} are thus

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}.$$

The chi-squared test of independence

Under H_0, the MLEs for the π_{ij} are thus

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}.$$

Under H_A, the MLEs are simply

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}.$$

The chi-squared test of independence

These MLEs can be used to form an LRT or the asymptotically equivalent Pearson's chi-squared test

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$E_{ij} = n\hat{\pi}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

are the counts expected under the null, giving (again!)

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}}.$$

The chi-squared test of independence

For the degrees of freedom, we see that

- under H_0, there are $(I-1)+(J-1)$ free parameters,
- under H_A, there are and $IJ-1$ free parameters.

The chi-squared test of independence

For the degrees of freedom, we see that

- under H_0, there are $(I-1)+(J-1)$ free parameters,
- under H_A, there are and $IJ-1$ free parameters.

So the asymptotic χ^2 distribution has (again!)

$$(IJ-1)-((I-1)+(J-1))=IJ-I-J+1=(I-1)(J-1)$$

degrees of freedom.

The chi-squared test of independence

For the degrees of freedom, we see that

- under H_0, there are $(I-1)+(J-1)$ free parameters,
- under H_A, there are and $IJ-1$ free parameters.

So the asymptotic χ^2 distribution has (again!)

$$(IJ-1)-((I-1)+(J-1))=IJ-I-J+1=(I-1)(J-1)$$

degrees of freedom.

This is the **chi-squared statistics for independence**.

The chi-squared test of independence

Note that the chi-squared statistics for homogeneity and independence are identical in form and degrees of freedom: however, the underlying hypotheses and sampling schemes are different.

(Consider performing bootstrap variants of the tests instead.)

Example: Marriage

For the marriage data,

```
R> chisq.test(tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 15.405, df = 1, p-value = 8.675e-05
```

Note that to reproduce the results in Rice one needs

```
R> chisq.test(tab, correct = FALSE)

        Pearson's Chi-squared test

data:  tab
X-squared = 16.01, df = 1, p-value = 6.302e-05
```

Outline

- Comparing two samples
  - Methods based on the normal distribution
  - A nonparametric method: The Mann-Whitney test
  - Methods based on the normal distribution
  - A nonparametric method: The signed rank test
- Analysis of categorical data
  - Fisher's exact test
  - The chi-squared test of homogeneity
  - The chi-squared test of independence
  - Matched-pairs designs

Motivation

Does the use of cell phones while driving cause accidents?

This is hard to study empirically (if usage is hazardous, it would be unethical to deliberately expose drivers to risk, etc.).

Redelmaier and Tibshirani (1997) conducted the following clever study.

699 drivers who owned cell phones and had been involved in motor vehicle collisions were identified. Then, billing records were used to determine whether each individual used a cell phone during the 10 minutes preceding the collision and also at the same time during the previous week. Hence, each person serves as its own control.

Motivation

Results were as follows:

	Before Collision		
At Collision	*On Phone*	*Not On Phone*	*Total*
On Phone	*13*	*157*	*170*
Not On Phone	*24*	*505*	*529*
Total	*37*	*662*	*699*

```
R> tab <- matrix(c(13, 157, 24, 505), nrow = 2, byrow = TRUE)
R> dimnames(tab) <- list(c("A_y", "A_n"), c("B_y", "B_n"))
R> tab

    B_y B_n
A_y  13 157
A_n  24 505
```

Motivation

A visual comparison of the frequencies:

```
R> mosaicplot(t(tab), main = "")
```

McNemar's test

We can model the data as a sample of size 699 from a multinomial distribution with four cells and respective cell probabilities π_{ij}.

The null hypothesis is that of marginal symmetry (distributions are the same at collision and before collision):

$$\pi_{11} + \pi_{21} = \pi_{\cdot 1} = \pi_{1\cdot} = \pi_{11} + \pi_{12}, \qquad \pi_{12} + \pi_{22} = \pi_{\cdot 2} = \pi_{2\cdot} = \pi_{21} + \pi_{22},$$

or equivalently,

$$H_0 : \pi_{12} = \pi_{21}.$$

McNemar's test

The MLEs of the relevant cell probabilities under H_0 are

$$\hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n}.$$

Under H_A, as before,

$$\hat{\pi}_{12} = \frac{n_{12}}{n}, \qquad \hat{\pi}_{21} = \frac{n_{21}}{n}$$

McNemar's test

These MLEs can be used to form an LRT or the asymptotically equivalent Pearson's chi-squared test

$$X^2 = \frac{(n_{12} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} + \frac{(n_{21} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

with $2 - 1 = 1$ (or $3 - 2 = 1$) degree of freedom.

McNemar's test

These MLEs can be used to form an LRT or the asymptotically equivalent Pearson's chi-squared test

$$X^2 = \frac{(n_{12} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} + \frac{(n_{21} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

with $2 - 1 = 1$ (or $3 - 2 = 1$) degree of freedom.

This is the **McNemar test**.

Example: Phones

By hand,

```
R> n12 <- tab[1, 2]
R> n21 <- tab[2, 1]
R> Xsq <- (n12 - n21)^2 / (n12 + n21)
R> Xsq
```

```
[1] 97.72928
```

Using a built-in classical test function:

```
R> mcnemar.test(tab)
```

```
        McNemar's Chi-squared test with continuity correction

data:  tab
McNemar's chi-squared = 96.265, df = 1, p-value < 2.2e-16
```

Outline

- Comparing two samples
- Analysis of categorical data
- Summary

Summary

Observe the correspondences of the classical test problems for metric and categorical data:

Summary

Observe the correspondences of the classical test problems for metric and categorical data:

- Testing the null $F_1 = \cdots = F_K$ that the distributions of J independent samples are the same: the **K-sample problem**. We covered $K = 2$ for numeric variables (t test for independent samples; Mann-Whitney aka Wilcoxon rank sum test) and the general case (e.g., chi-squared test for homogeneity) for categorical data.

Summary

Observe the correspondences of the classical test problems for metric and categorical data:

- Testing the null $F_1 = \cdots = F_K$ that the distributions of J independent samples are the same: the K**-sample problem**. We covered $K = 2$ for numeric variables (t test for independent samples; Mann-Whitney aka Wilcoxon rank sum test) and the general case (e.g., chi-squared test for homogeneity) for categorical data.
- For pairs (X, Y) of observations, test the null of independence of X and Y, the so-called **independence problem** (or **contingency problem**). We covered only the categorical case (e.g., chi-squared test for independence); see e.g. `cor.test` for variants for numeric data.

Summary

- For pairs (X, Y) of (not necessarily independent) observations, test the null $F_X = F_Y$: the **symmetry problem**. We covered both the numeric (t test for paired samples; (Wilcoxon) signed rank test) and the categorical case (McNemar test).

Summary

- For pairs (X, Y) of (not necessarily independent) observations, test the null $F_X = F_Y$: the **symmetry problem**. We covered both the numeric (t test for paired samples; (Wilcoxon) signed rank test) and the categorical case (McNemar test).

Observe also that in many cases, there are modern conditional (permutation based) tests as (preferable) alternatives to the classical unconditional tests.