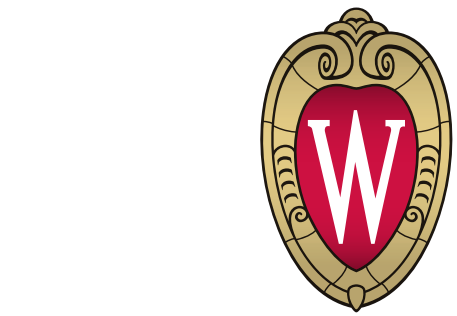
Chapter 8: Comparing two independent populations

Part 2

https://dzwang91.github.io/stat324/



WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Outline

What is permutation?

- Permutation refers to the arrangement of a set of objects into some specified order.

What is permutation?

- Permutation refers to the arrangement of a set of objects into some specified order.

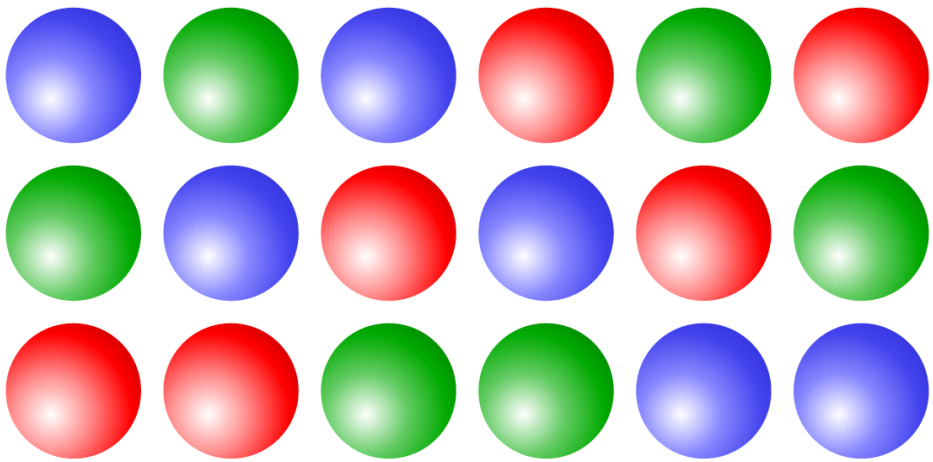


Figure: Each column is one possible permutation of three columns.

Permuting a data set

- Given a data set of sample size $n = 3$: X_1, X_2, X_3, there are 6 possible permutations:
  - $X_{(1)} = (X_1, X_2, X_3)$
  - $X_{(2)} = (X_1, X_3, X_2)$
  - $X_{(3)} = (X_2, X_1, X_3)$
  - $X_{(4)} = (X_2, X_3, X_1)$
  - $X_{(5)} = (X_3, X_1, X_2)$
  - $X_{(6)} = (X_3, X_2, X_1)$

Permuting a data set

- Given a data set of sample size $n = 3$: X_1, X_2, X_3, there are 6 possible permutations:
  - $X_{(1)} = (X_1, X_2, X_3)$
  - $X_{(2)} = (X_1, X_3, X_2)$
  - $X_{(3)} = (X_2, X_1, X_3)$
  - $X_{(4)} = (X_2, X_3, X_1)$
  - $X_{(5)} = (X_3, X_1, X_2)$
  - $X_{(6)} = (X_3, X_2, X_1)$
- In general, there are $n!$ permutations for a data set with sample size n

Why are permutations useful?

- Hypothesis testing paradigm:
  - Collect some data
  - Form the null hypothesis
  - Design the test statistic
  - Derive sampling distribution of test statistic under H_0

Why are permutations useful?

- Hypothesis testing paradigm:
  - Collect some data
  - Form the null hypothesis
  - Design the test statistic
  - Derive sampling distribution of test statistic under H_0
- In many cases, the null hypothesis is nil hypothesis, i.e., no effect
- Under H_0, all permutations are equally likely, so permutations relate to sampling distribution

Permutation test

- Two independent samples from two populations, label them 1 and 2
  - $(X_1, ..., X_m)$: sample from population 1
  - $(Y_1, ..., Y_n)$: sample from population 2
  - μ_1: true mean in population 1
  - μ_2: true mean in population 2
  - $\bar{X}$: sample mean of the first sample
  - $\bar{Y}$: sample mean of the second sample
- We wish to test: $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$.

Permutation test

- Two independent samples from two populations, label them 1 and 2
  - $(X_1, ..., X_m)$: sample from population 1
  - $(Y_1, ..., Y_n)$: sample from population 2
  - μ_1: true mean in population 1
  - μ_2: true mean in population 2
  - $\bar{X}$: sample mean of the first sample
  - $\bar{Y}$: sample mean of the second sample
- We wish to test: $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$.
- Test statistic: $T(X_1, ..., X_m, Y_1, ..., Y_n) = \bar{X} - \bar{Y}$

Permutation test

Given two original data set: $X_1, X_2, ..., X_m$ and $Y_1, ..., Y_n$.

1. Compute the sample mean $\bar{X}$ and $\bar{Y}$, and the realization of the test statistic:

$$T_{obs} = \bar{X} - \bar{Y}$$

Permutation test

Given two original data set: $X_1, X_2, ..., X_m$ and $Y_1, ..., Y_n$.

1. Compute the sample mean $\bar{X}$ and $\bar{Y}$, and the realization of the test statistic:

$$T_{obs} = \bar{X} - \bar{Y}$$

One run of permuation

2. Combine all data from both samples into a single group of size $m + n$. Take a random sample of size m, without replacement, from this group, and compute its mean, call it $\bar{X}^*$. Compute the mean of the remaining data points, call it $\bar{Y}^*$.
3. Compute a new realization of statistic $T^*_{obs} = \bar{X}^* - \bar{Y}^*$.

Permutation test

Given two original data set: $X_1, X_2, ..., X_m$ and $Y_1, ..., Y_n$.

1. Compute the sample mean $\bar{X}$ and $\bar{Y}$, and the realization of the test statistic:

$$T_{obs} = \bar{X} - \bar{Y}$$

One run of permuation

2. Combine all data from both samples into a single group of size $m + n$. Take a random sample of size m, without replacement, from this group, and compute its mean, call it $\bar{X}^*$. Compute the mean of the remaining data points, call it $\bar{Y}^*$.
3. Compute a new realization of statistic $T^*_{obs} = \bar{X}^* - \bar{Y}^*$.

4. Repeat steps 2-3 B times, and compute T^*_{obs} from each one. Let m be the number of values of T^*_{obs} that are less than or equal to $-|T_{obs}|$ or greater than or equal to $|T_{obs}|$. Then p-value is $\frac{m}{B}$.

Permutation test in R

```
> rm(list=ls())
>
> ## data set
> dead <- c(17.65, 20.83, 24.59, 18.52, 21.40, 23.78, 20.36, 18.83, 21.83, 20.06)
> live <- c(23.76, 21.17, 26.13, 20.18, 23.01, 24.84, 19.34, 24.94, 27.14, 25.87,
18.95, 22.61)
> all <- c(dead, live)
>
> m=length(dead)
> n=length(live)
>
> # calculate the observation of test statistic
>
> tobs=mean(dead)-mean(live)
```

Figure: imput data sets, merge into a single one, calculate the observation of test statistic

Permutation test in R

```
> # build a permutation test function
> permutationtest=function(data,obs,b) {
+   permutestat=NULL
+   m=0
+   for(i in 1:b) {
+     permutationdata=all[sample.int(m+n)]
+     firstsamp=permutationdata[1:m]
+     secondsamp=permutationdata[(m+1):(m+n)]
+     firstmean=mean(firstsamp)
+     secondmean=mean(secondsamp)
+     permutestat[i]=firstmean-secondmean
+     if (isTRUE(permutestat[i]>=abs(obs)|permutestat[i]<=-abs(obs))){
+       m=m+1
+     } else {
+       m=m
+     }
+   }
+   pvalue=m/b
+   return(pvalue)
+ }
```

Figure: make a permutation test function

Permutation test in R

```
> # run b times
> b = 10000
> permutationtest(all,tobs,b)
[1] 0.0011
```

Figure: run permutation test for b times

Outline

Example

- "Does handedness differ according to sex?"

Example

- "Does handedness differ according to sex?"
- Let π_{FL} be the proportion of females that are left-handed, and π_{ML} be the proportion of males that are left-handed, then we want to test:

$$H_0 : \pi_{FL} - \pi_{ML} = 0$$
$$H_A : \pi_{FL} - \pi_{ML} \neq 0$$

Example

- "Does handedness differ according to sex?"
- Let π_{FL} be the proportion of females that are left-handed, and π_{ML} be the proportion of males that are left-handed, then we want to test:

$$H_0 : \pi_{FL} - \pi_{ML} = 0$$
$$H_A : \pi_{FL} - \pi_{ML} \neq 0$$

- A sample of $n_M = 54$ males and $n_F = 21$ females was taken, and each person was asked to indicate which was their dominant hand.
- The data are as follows:

Female: 12 left, 9 right
Male: 23 left, 31 right

Two independent population proportions

- Two separate samples from two populations, label them 1 and 2.
  - π_1 = true proportion in population 1
  - π_2 = true proportion in population 2
  - n_1 = sample size taken from population 1
  - n_2 = sample size taken from population 2
  - P_1 = sample proportion from first sample
  - P_2 = sample proportion from second sample
- We wish to test: $H_0 : \pi_1 - \pi_2 = 0$ vs. $H_A : \pi_1 - \pi_2 \neq 0$.

Two independent population proportions

- Assume $\pi_1 n_1, (1-\pi_1)n_1, \pi_2 n_2, (1-\pi_2)n_2$ are all greater than 5:

$$P_1 \approx N(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1})$$

$$P_2 \approx N(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2})$$

Two independent population proportions

- Assume $\pi_1 n_1, (1-\pi_1)n_1, \pi_2 n_2, (1-\pi_2)n_2$ are all greater than 5:

$$P_1 \approx N(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1})$$

$$P_2 \approx N(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2})$$

- Assume all of the data points are independent, both within and between populations:

$$P_1 - P_2 \approx N(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2})$$

Two independent population proportions

- Therefore under H_0: $\pi_1 = \pi_2 = \pi$, we have

$$P_1 - P_2 \approx N(0,\ \pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2}))$$

Two independent population proportions

- Therefore under H_0: $\pi_1 = \pi_2 = \pi$, we have

$$P_1 - P_2 \approx N(0,\ \pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2}))$$

Equivalently,

$$\frac{P_1 - P_2}{\sqrt{\pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0, 1)$$

Two independent population proportions

- Therefore under H_0: $\pi_1 = \pi_2 = \pi$, we have

$$P_1 - P_2 \approx N(0,\ \pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2}))$$

Equivalently,

$$\frac{P_1 - P_2}{\sqrt{\pi(1-\pi)(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0, 1)$$

- But π is unknown,

Two independent population proportions

- Therefore under H_0: $\pi_1 = \pi_2 = \pi$, we have

$$P_1 - P_2 \approx N(0,\ \pi(1-\pi)(\tfrac{1}{n_1} + \tfrac{1}{n_2}))$$

  Equivalently,

$$\frac{P_1 - P_2}{\sqrt{\pi(1-\pi)(\frac{1}{n_1}+\frac{1}{n_2})}} \approx N(0,1)$$

- But π is unknown, so we estimate π using a weighted average of the two individual sample proportions:

$$P = \frac{P_1 n_1 + P_2 n_2}{n_1 + n_2}.$$

Two independent population proportions

- Therefore under H_0: $\pi_1 = \pi_2 = \pi$, we have

$$P_1 - P_2 \approx N(0,\ \pi(1-\pi)(\tfrac{1}{n_1} + \tfrac{1}{n_2}))$$

Equivalently,

$$\frac{P_1 - P_2}{\sqrt{\pi(1-\pi)(\frac{1}{n_1}+\frac{1}{n_2})}} \approx N(0,1)$$

- But π is unknown, so we estimate π using a weighted average of the two individual sample proportions:

$$P = \frac{P_1 n_1 + P_2 n_2}{n_1 + n_2}.$$

- Then the test statistic is:

$$Z = \frac{P_1 - P_2}{\sqrt{P(1-P)(\frac{1}{n_1}+\frac{1}{n_2})}}.$$

Recap

- We need to check:
  - All of the data points are independent, both within and between populations
  - The sample sizes are large enough (πn_1, $(1-\pi)n_1$, πn_2, and $(1-\pi)n_2$ are all greater than 5)
- Under H_0, the test statistic

$$Z = \frac{P_1 - P_2}{\sqrt{P(1-P)(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1)$$

- Calculate the p-value or rejection region based on the given significance level α to make a conclusion

Back to example

- $P_{FL} = 0.571$, $P_{ML} = 0.426$, and $P_L = \frac{12+23}{21+54} = 0.467$.
- The realization of test statistic:

$$Z_{obs} = \frac{0.571-0.426-0}{\sqrt{0.467(1-0.467)(\frac{1}{21}+\frac{1}{54})}} = 1.13.$$

- p-value $= 2 \times P(Z > 1.13) = 0.258$. If $\alpha = 0.05$, we conclude that there is not enough evidence to say that males and females have a different proportion of left-handed individuals.

What's the next?

We'll discuss how to compare two paired populations in next lecture.