ATLAS
ASIAN CLINICAL TRIALS NETWORK FOR CANCERS PROJECT

National Cancer Center Japan

Hypothesis Testing 2

National Cancer Center, Japan

Biostatistics Division

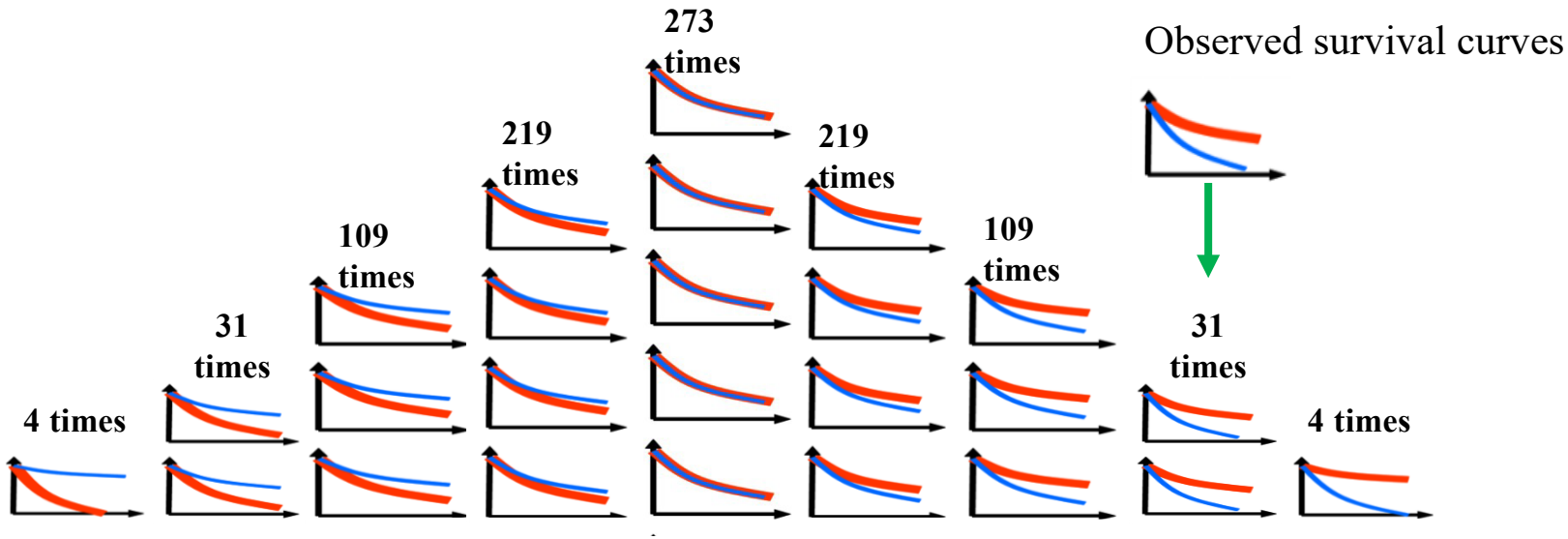
Aya Kuchiba

JCOG
Japan Clinical Oncology Group

Summary of the Previous Lecture

Comparison of Survival Curves between RT and CRT groups

- Distribution under the null hypothesis “no difference between RT and CRT”
- The probability of the difference being greater than the observed difference between RT and CRT = *p*-value



Summary of the Previous Lecture

Definition and Interpretation of the p-value

- *p*-value is the value of the probability of something more extreme occurring than the data under the null hypothesis "no difference between RT and CRT“
- Large *p*-value
  → The obvious happened
  → Accept the null hypothesis "no difference"
- Small *p*-value
  → Rare occurrence
  → Reject the null hypothesis of "no difference" and conclude that there is a difference.

Today’s Goal: To Understand Various Testing Methods

Various types of tests

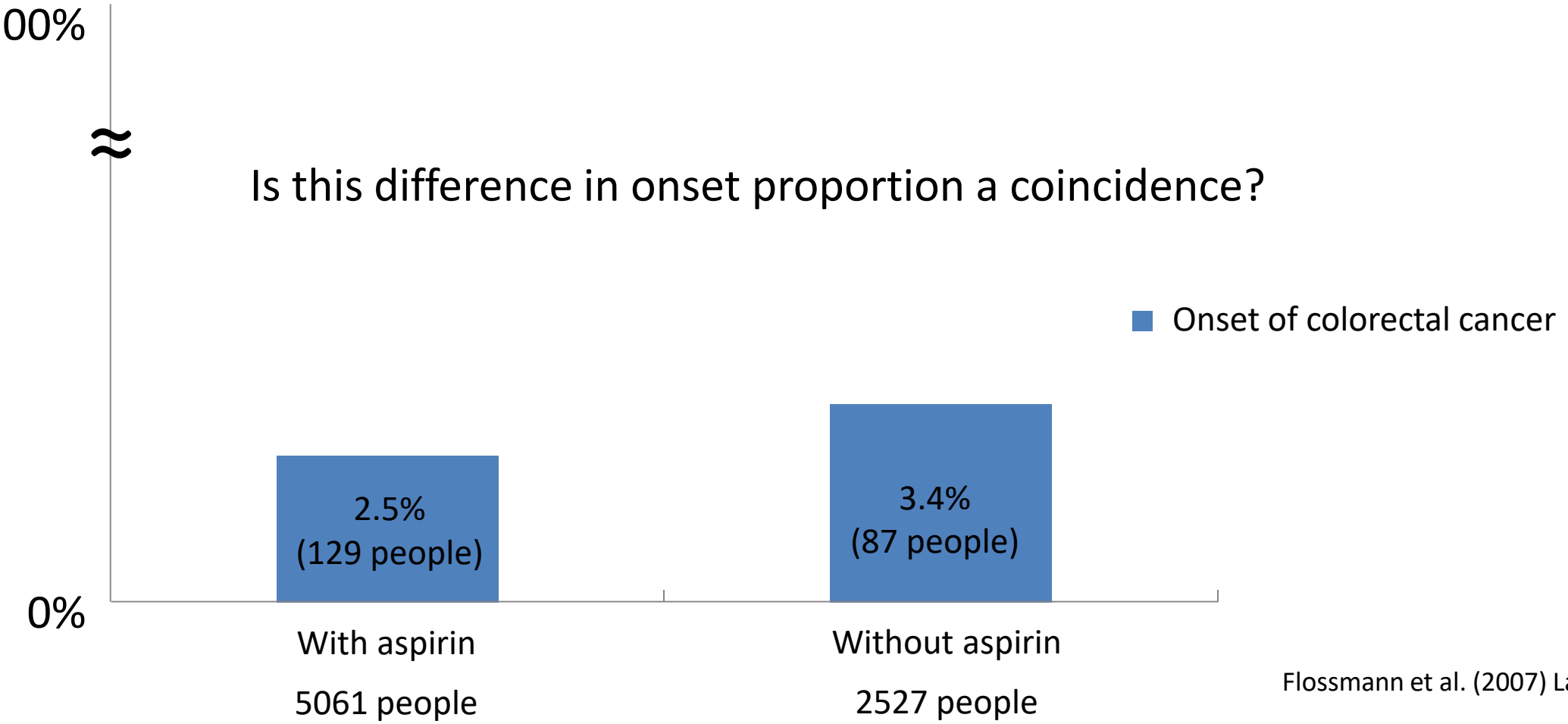
- Data type
  - Binary values, such as onset “yes” and “no”
  - Continuous variables, such as blood pressure and body mass index (BMI)
  - Discrete variables, such as quality-of-life (QOL) score “0, 1, 2, 3”
- Study design
  - How the data was collected
- Hypothesis to be examined

Today's Lecture

- Binary data
  - Data summary: 2×2 contingency tables
  - Fisher's exact test
  - Chi-square test
- Continuous variable data
  - Data summary: Center of distribution and variability
  - t-test
  - Wilcoxon rank sum test
- Discrete variable data
- Summary

Example of Binary Data

Research question: Does aspirin prevent colorectal cancer?



Flossmann et al. (2007) Lancet

Binary Data Summary: 2×2 Contingency Table

- Outcomes of 2 groups × 2 values

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129 people	4932 people	5061 people
	No	87 people	2440 people	2527 people
	Total	216 people	7372 people	7588 people

For Data Obtained from Randomized Controlled Trials

- Research question: Does aspirin prevent colorectal cancer?
- Randomized controlled trial

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129	4932	5061
	No	87	2440	2527
	Total	216	7372	7588

Flossmann et al. (2007) Lancet

The Number of Participants in Each Group is Fixed According to the Study Design

- Researchers decide

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129	4932	5061
	No	87	2440	2527
	Total	216	7372	7588

Colorectal Cancer Onset Outcomes Can Also Be Fixed Under the Null Hypothesis

If there were truly no association between aspirin and the onset of colorectal cancer,

- the 216 people who developed colorectal cancer would develop the disease with or without aspirin.
- the 7372 people who did not develop colorectal cancer did not develop the disease with or without aspirin.

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129	4932	5061
	No	87	2440	2527
	Total	216	7372	7588

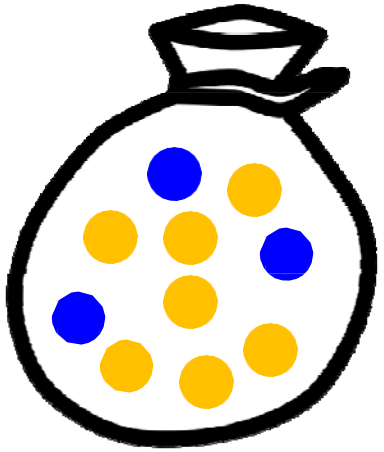
Marginal Totals Are Fixed Under the Null Hypothesis

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129	4932	5061
	No	87	2440	2527
	Total	216	7372	7588

- Under the null hypothesis, randomization is the same thing as dividing 7588 people consisting of 216 people who develop the disease and 7372 people who do not develop the disease into 5061 and 2527 people randomly.
- Probability calculations based on hypergeometric distributions

Example of Hypergeometric Distribution

- If you take 5 balls out of a bag that contains 3 blue balls and 7 orange balls, what is the probability that **1** of them is a blue ball?



	Blue	Orange	Total
Balls taken from the bag	**1**	4	5
Balls left in the bag	2	3	5
Total	3	7	10

Probability of obtaining a contingency table for this pattern = $\frac{{}_3C_1 \, {}_7C_4}{{}_{10}C_5}$

Fisher’s Exact Test

- Under fixed marginal totals, determine all the resulting contingency table patterns and the probability of each occurring, and calculate the probability of extreme results over and above the observed contingency table.

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	a	b	5061
	No	c	d	2527
	Total	216	7372	7588

Fisher's Exact Test

- Consider all patterns

216	4845	5061
0	2527	2527
216	7372	7588

Pr = 2.4 × 10^{-37}

215	4846	5061
1	2526	2527
216	7372	7588

Pr = 1.3 × 10^{-35}

214	4847	5061
2	2525	2527
216	7372	7588

Pr = 5.0 × 10^{-34}

. . .

129	4932	5061
87	2440	2527
216	7372	7588

Pr = 0.0053

. . .

0	5061	5061
216	2311	2527
216	7372	7588

Pr = 1.2 × 10^{-106}

The same contingency table as the observed data

- One-sided *P* value = 0.0053 + … + 1.2 × 10^{-106} = 0.0174

(Significant at the 2.5% level on one side!)

In a Randomized Controlled Trial, the Test Method Is Determined Based on "Randomization."

Data type	Test methods
Binary data	Fisher’s exact test
Continuous variable data	Sorting test Wilcoxon rank sum test, etc.
Survival time data	Log-rank test, etc.

Can We Use Fisher's Exact Test Without Randomization?

- Fixing the marginal total can be justified in many nonrandomized studies.
- For example, observational studies with cohort study designs
  - Groups receiving or not receiving aspirin can be fixed as a part of study design.
  - Under the null hypothesis, the presence or absence of disease onset can also be fixed.
  - Fisher's Exact test can still be used.

Today's Lecture

- Binary data
  - Data summary: 2×2 contingency tables
  - Fisher's exact test
  - Chi-square test
- Continuous variable data
  - Data summary: Center of distribution and variability
  - t-test
  - Wilcoxon rank sum test
- Discrete variable data
- Summary

For Data Obtained by Random Sampling from a Population

- Research question: Does aspirin prevent colorectal cancer?
- Random sampling from population
  - Randomly sampled 5061 participants from the group who usually take aspirin and 2527 participants from the group who do not usually take aspirin and followed the onset of the disease.

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129	4932	5061
	No	87	2440	2527
	Total	216	7372	7588

Equivalent to an observational study where the participants could be randomly sampled

Expected Contingency Table under the Null Hypothesis

- If aspirin was not effective, the onset proportion rate for both groups should be 216/7588 ≈ 2.8%.
- Expected contingency table when aspirin is ineffective

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129 (2.8%)	4932 (97.2%)	5061
	No	87 (2.8%)	2440 (97.2%)	2527
	Total	216 (2.8%)	7372 (97.2%)	7588

- 5061 × 216/7588 = 144.06 ≈ 144
- 2527 × 216/7588 = 71.93 ≈ 72

Expected and Observed Contingency Tables

- Expected contingency table under the null hypothesis

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129 (2.8%)	4932 (97.2%)	5061
	No	87 (2.8%)	2440 (97.2%)	2527
	Total	216 (2.8%)	7372 (97.2%)	7588

- Observed contingency table

		Onset of colorectal cancer		
		Yes	No	Total
Aspirin	Yes	129 (2.5%)	4932 (97.5%)	5061
	No	87 (3.4%)	2440 (96.6%)	2527
	Total	216 (2.8%)	7372 (97.2%)	7588

Chi-square Test

Observed values		Expected values	
129 (2.5%)	4932 (97.5%)	144 (2.8%)	4917 (97.2%)
87 (3.4%)	2440 (96.6%)	72 (2.8%)	2455 (97.2%)

- Test based on the difference between observed and expected frequencies
  - The greater the difference between the observed and expected frequencies, the rarer the result observed under the null hypothesis.
- Statistic $\frac{(1239 - 144)^2}{144} + \frac{(4932 - 4917)^2}{4917} + \frac{(87 - 72)^2}{72} + \frac{(2440 - 2455)^2}{2455}$
- Approximates a chi-square distribution with one degree of freedom
  - The higher the expected frequency, the better the approximation.
- P = 0.0273 (Significant at the two-sided 5% level!)

Fisher’s Exact Test and Chi-Square Test

- Fisher’s exact test
  - Fixing the marginal total
  - Exact probability distribution
  - Two-sided P = 0.0334
- Chi-square test
  - Random sampling and sufficient sample size
  - Approximation to chi-square distribution
  - Two-sided P = 0.0273

- The conditions for the chi-square test are difficult to achieve, as it is rarely possible to perform random sampling.
- In many cases, it is correct to use Fisher’s exact test.
- However, if the sample size is large, the results are almost identical,

In most cases, the test does not matter
(The result remains the same)

Testing using 2 × 2 Contingency Tables

		Onset of colorectal cancer	
		Yes	No
Aspirin	Yes	2	6
	No	4	4

- Fisher's exact test if the sample size is small and chi-square test if sample size is large
- Let us perform the Yates's chi-square test with continuity correction

Testing using 2 × 2 Contingency Tables

- Should Fisher's exact test be used for a small sample size in a 2 × 2 contingency table, while the chi-square test is appropriate for a larger sample size?
- Yates's chi-square test with continuity correction for a 2 × 2 contingency table!

- Fisher's Exact Test if it is a randomization test
- Fisher's exact test is also suitable in many observational studies
  - The chi-square test can be considered an approximation of Fisher's exact test and is therefore acceptable for large sample sizes.
- Yates's continuous correction is an approximation of Fisher's exact test
  - We can use the Fisher's exact test

Today's Lecture

- Binary data
  - Data summary: 2×2 contingency tables
  - Fisher's exact test
  - Chi-square test
- Continuous variable data
  - Data summary: Center of distribution and variability
  - t-test
  - Wilcoxon rank sum test
- Discrete variable data
- Summary

Characteristics of Mean

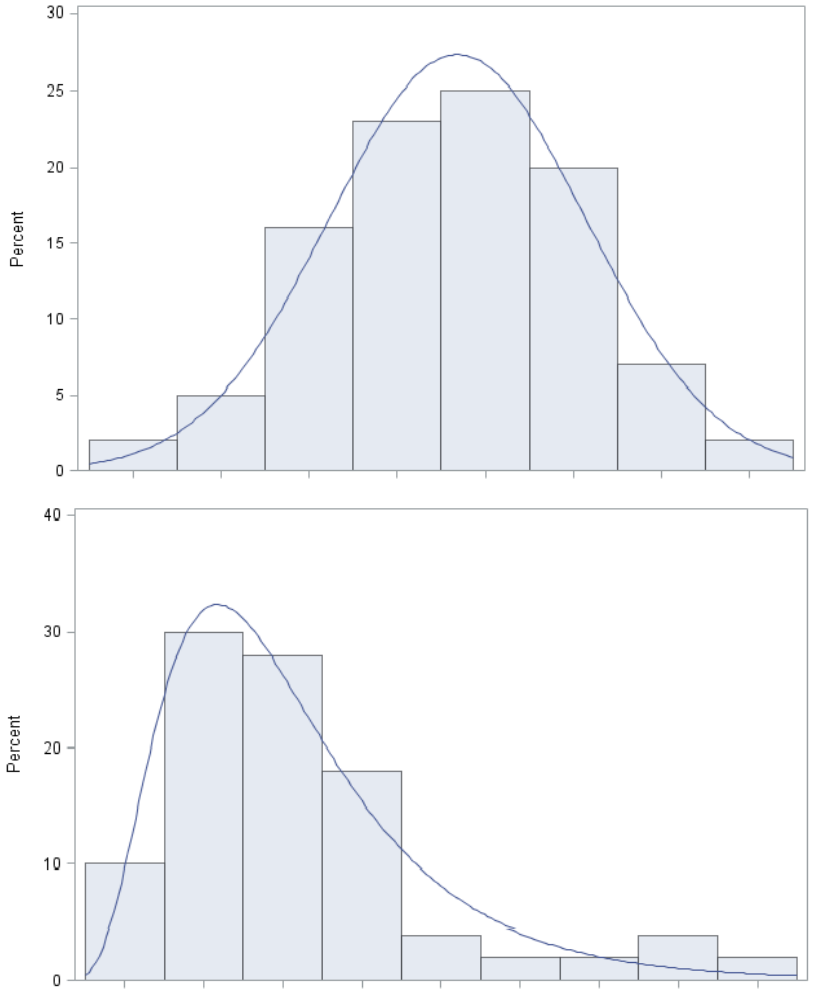
- Numerical example

Data	Mean	Median
1, 3, 5, 7, 9	5	5
1, 3, 5, 7, **100**	23.2	5

  - Mean: all values added together and divided by the number of values
  - Median: middle value of the dataset
- Means are more sensitive to extreme values than median.
  - Extreme values = Outliers

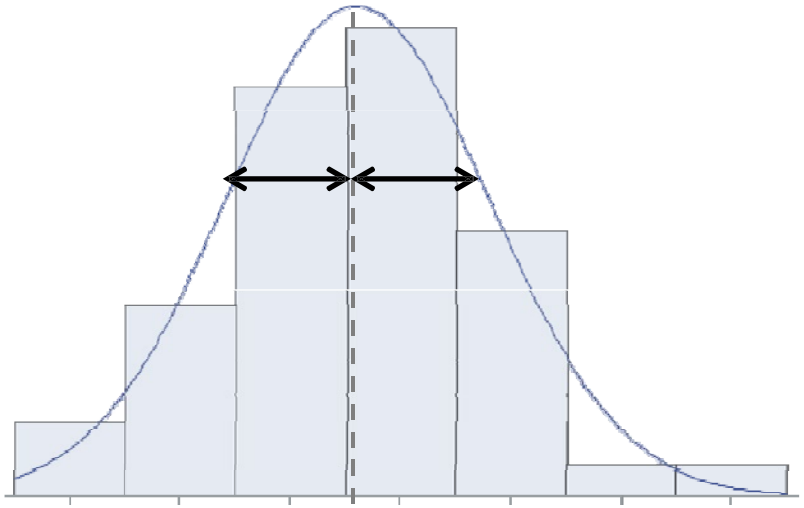
Data Summary: Center of Distribution

- Mean
  - If the histogram is approximately bilaterally symmetric
  - There is no outlier
  - The mean and the median are approximately the same
- Median
  - Other than above
  - For example, the histogram is distorted



Data Summary: Variation in Distribution

- Standard Deviation (SD)
  - How scattered the data are around the mean?
  - Variance = (standard deviation)^2
- Minimum, Maximum
- Quantile Points
  - For example, the 25th and 75th percentile points



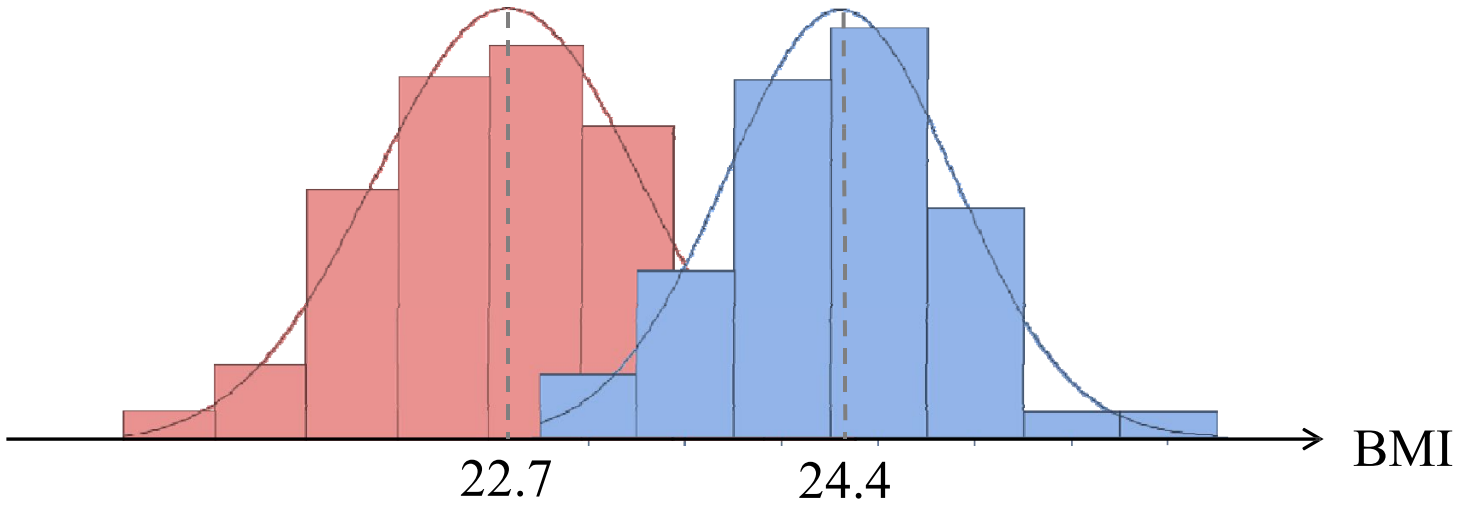
Comparison of Continuous Data

- Research question: Do diet method A (group A) and diet method B (group B) differ in post-diet BMI?
- Null hypothesis: post-diet BMI is equal for groups A and B.

BMI (Kg/m^2)			
Group A (n = 100)	22.0	· · ·	22.3
Group B (n = 100)	23.3	· · ·	23.5

Check the Shape of the Distribution using a Histogram

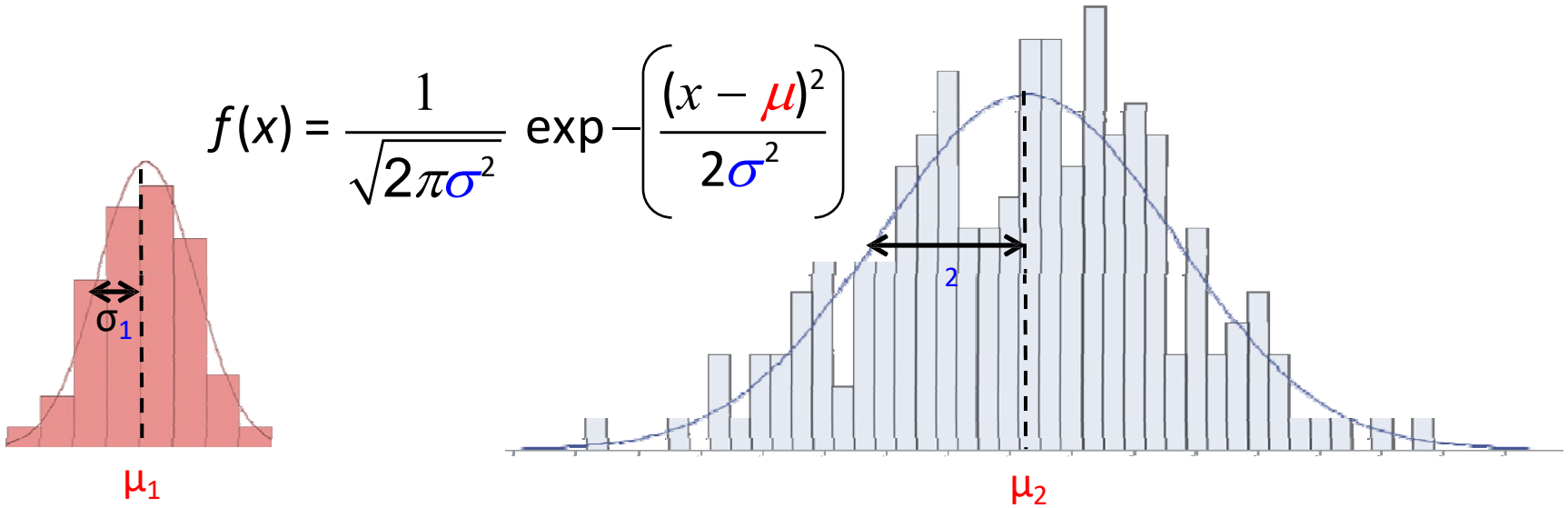
BMI (Kg/m^2)				Mean
Group A (n = 100)	22.0	· · ·	22.3	22.7
Group B (n = 100)	23.3	· · ·	23.5	24.4



Normal Distribution

- Mean and variance determine the shape of the distribution.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\left(\frac{(x-\mu)^2}{2\sigma^2} \right)$$



- If the data follow a normal distribution, and it can be assumed that the variances of the two groups are equal, then "comparison of means" is equivalent to "comparison of distributions."

Comparing the Means of Two Groups: *t*-test

- *t* statistic

$$t = \frac{\overline{X_A} - \overline{X_B}}{\sqrt{V\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$\overline{X_A}$: Mean in group A
$\overline{X_B}$: Mean in group B
V: Estimated variance common to both groups
n_A: Group A sample size
n_B: Group B sample size

- When the data follow a normal distribution, the t statistic follows a t distribution.
- Calculate the t statistic from the data and calculate the P value by comparing the t statistic to the t distribution.
  - Based on the P value, determine whether there is a difference in the effect of the diet method.

Testing Continuous Data

BMI (kg/m^2)				
Group A (n = 100)	22.0	· · ·	· · ·	22.3
Group B (n = 100)	23.3	· · ·	· · ·	23.5

- If the variances differ between groups when testing for equal variances, perform a Welch's t-test
- Perform a normality test and if the data do not follow a normal distribution, do not perform a *t*-test

Welch's *t*-test

- Welch's *t*-statistic

$$t = \frac{\overline{X_A} - \overline{X_B}}{\sqrt{\left(\frac{V_A}{n_A} + \frac{V_B}{n_B}\right)}}$$

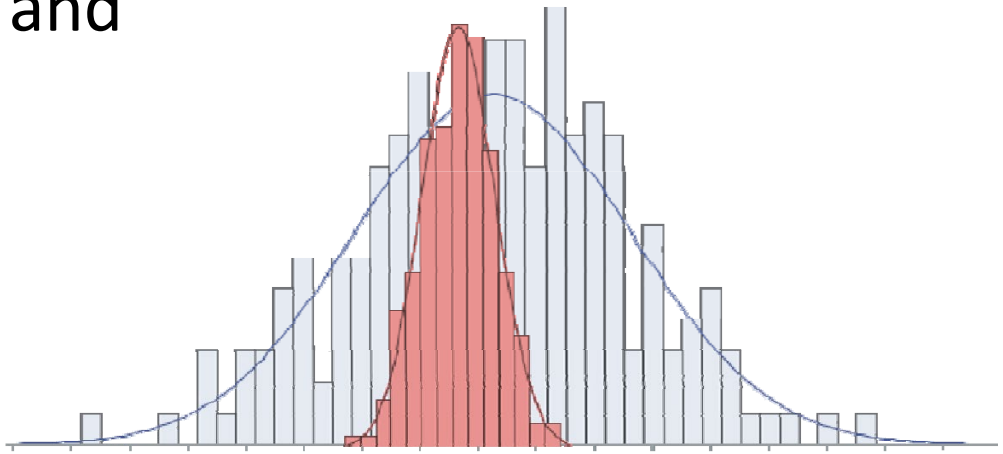
$\overline{X_A}$: Mean in group A
$\overline{X_B}$: Mean in group B
V_A: Estimated variance in group A
V_B: Estimated variance in group B
n_A: Group A sample size
n_B: Group B sample size

- Welch's *t*-statistic follows an approximate *t*-distribution even when the variances of the two groups differ
- The *t*-statistic is determined from the data and the *P*-value is obtained to determine if the means of the two groups are equal

This is a valid method, but is it always the best choice?

When Variances are Not Equal

- Normal distribution is determined by mean and variance
- Already different distributions
  - This in itself is significant
  - We need to consider whether we still want to compare the means
- The *t*-test is relatively robust to differences in variance between groups
  - When the sample sizes of both groups are equal, there is no need to be particularly concerned.



When Data Do Not Follow a Normal Distribution

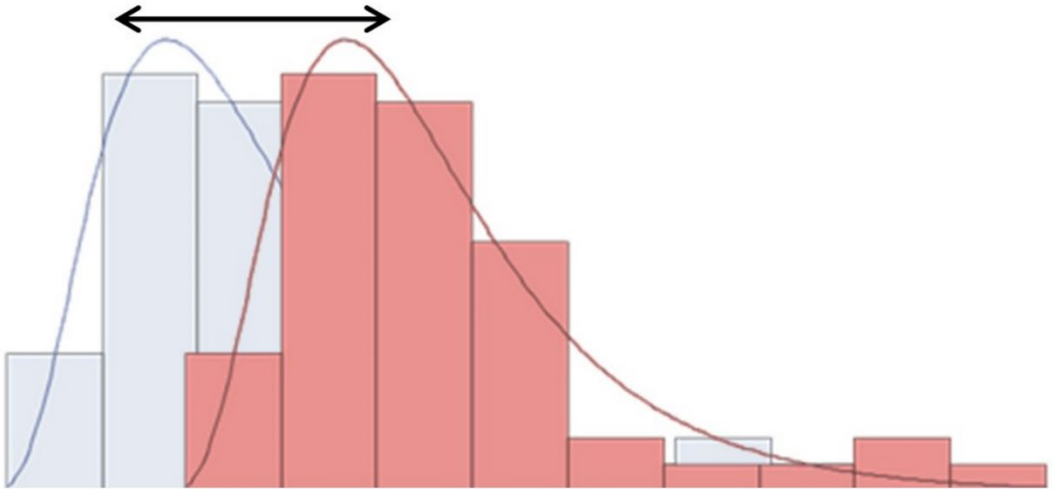
- The *t*-test is relatively robust to the normality assumption.
  - α error is kept below the nominal level.
  - However, too much deviation, skewness, or outliers will lower the power.

Comparing Continuous Data

- Draw a histogram to check the shape of the distribution
- If the shape of the distribution is not much different between the two groups,
  - perform a t-test
- When the shape of the distribution differs significantly among 2 groups,
  - consider whether it makes sense to compare means
  - If you want to compare, a t-test is fine in most cases

Another Way to Compare Continuous Data

- When the data deviate significantly from normal distribution, they may contain outliers.
- Nonparametric method
  - No distribution is assumed for the data
- Test for identifying difference in shape of distributions



Wilcoxon Rank Sum Test

	BMI (kg/m^2)				Mean
Group A	22.0 [2]	28.3 [8]	19.4 [1]	22.3 [3]	23.0
Group B	23.3 [4]	25.1 [7]	24.6 [6]	23.5 [5]	24.1

- Combine the groups and [rank] them in order of decreasing value
- Calculate the sum of the ranks for each group
  - Rank sum of group A = 2 + 8 + 1 + 3 = 14
  - Rank sum of group B = 4 + 7 + 6 + 5 = 22
- Not affected by outliers, as it is a "rank"
  - ‘Robust’ to outliers
  - 28.3 (Kg/m2) → 40.0 (kg/m^2)
    Mean of group A → 25.9
    Rank sum of group A → 14

Wilcoxon Rank Sum Test

- Consider the rank sum under the null hypothesis
  - If there was no difference between the two groups, the ranks should be equally distributed
  - (1 + 2 +...+ 8)/2 = 18
- Compare "observed rank sums" with "rank sums under the null hypothesis."
  - Observed value is 14
  - Under the null hypothesis, it is 18
  - Difference of 4
- Calculate p-value using theoretical distribution of the difference
  - P = 0.346(Not significant)
  - Conclude that the diet had no effect

Testing Continuous Data

BMI (kg/m^2)				
Group A (n = 4)	22.0	28.3	19.4	22.3
Group B (n = 4)	23.3	25.1	24.6	23.5

- Wilcoxon rank sum test if the sample size is small
  (*t*-test if the sample size is sufficient)

t-test and Wilcoxon Rank Sum Test

- *t*-test
  - Assuming a normal distribution, comparing means
    - Relatively robust to distributional assumptions
  - Under a normal distribution, it has the highest power

- Wilcoxon rank sum test
  - Without assuming a distribution, comparing rank sums
    - May be used under a normal distribution
    - The power is sufficiently high in this case
  - Robust to outliers

Today's Lecture

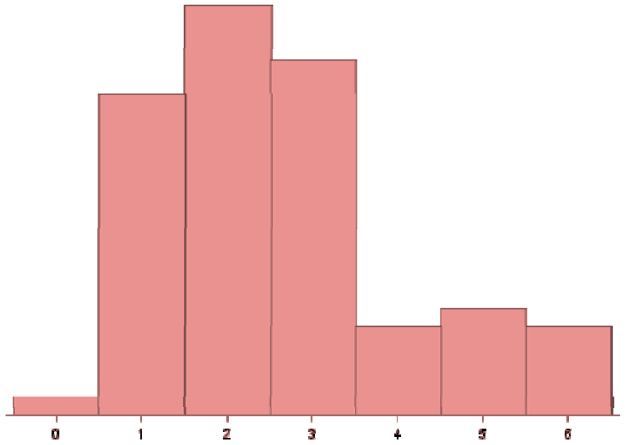
- Binary data
  - Data summary: 2×2 contingency tables
  - Fisher's exact test
  - Chi-square test
- Continuous variable data
  - Data summary: Center of distribution and variability
  - t-test
  - Wilcoxon rank sum test
- Discrete variable data
- Summary

Comparison of Discrete Variable Data

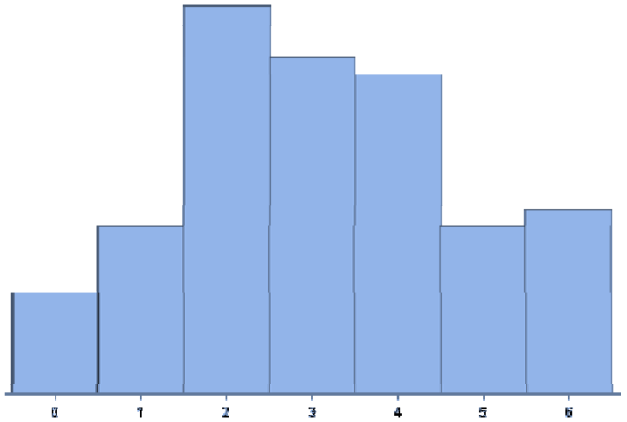
- QOL and satisfaction surveys, etc.

	very satisfied	satisfied	slightly satisfied	average	slightly unsatisfied	unsatisfied	very unsatisfied
score	0 point	1 point	2 points	3 points	4 points	5 points	6 points

- Draw a histogram of satisfaction scores between diet method groups A and B.



Group A



Group B

Comparing the Distribution of Discrete Variable Data

- *t*-test
  - Assuming a normal distribution, comparing means
    - Relatively robust to distributional assumptions
  - Under a normal distribution, it has the highest power
- Wilcoxon rank sum test
  - Without assuming a distribution, comparing rank sums
    - May be used under a normal distribution
    - The power is sufficiently high in this case
  - Robust to outliers

- Even for discrete data, *t*-tests and Wilcoxon rank sum tests can be used.

Today's Lecture

- Binary data
  - Data summary: 2×2 contingency tables
  - Fisher's exact test
  - Chi-square test
- Continuous variable data
  - Data summary: Center of distribution and variability
  - t-test
  - Wilcoxon rank sum test
- Discrete variable data
- Summary

Today’s Goal: To Understand Various Testing Methods

Various types of tests

- Data type
  - Binary values, such as onset “yes” and “no”
  - Continuous variables, such as blood pressure and BMI
  - Discrete variables, such as QOL score “0, 1, 2, 3”
- Study design
  - How the data was collected
- Hypothesis to be examined