HYPOTHESIS TESTING

ERT 474/574

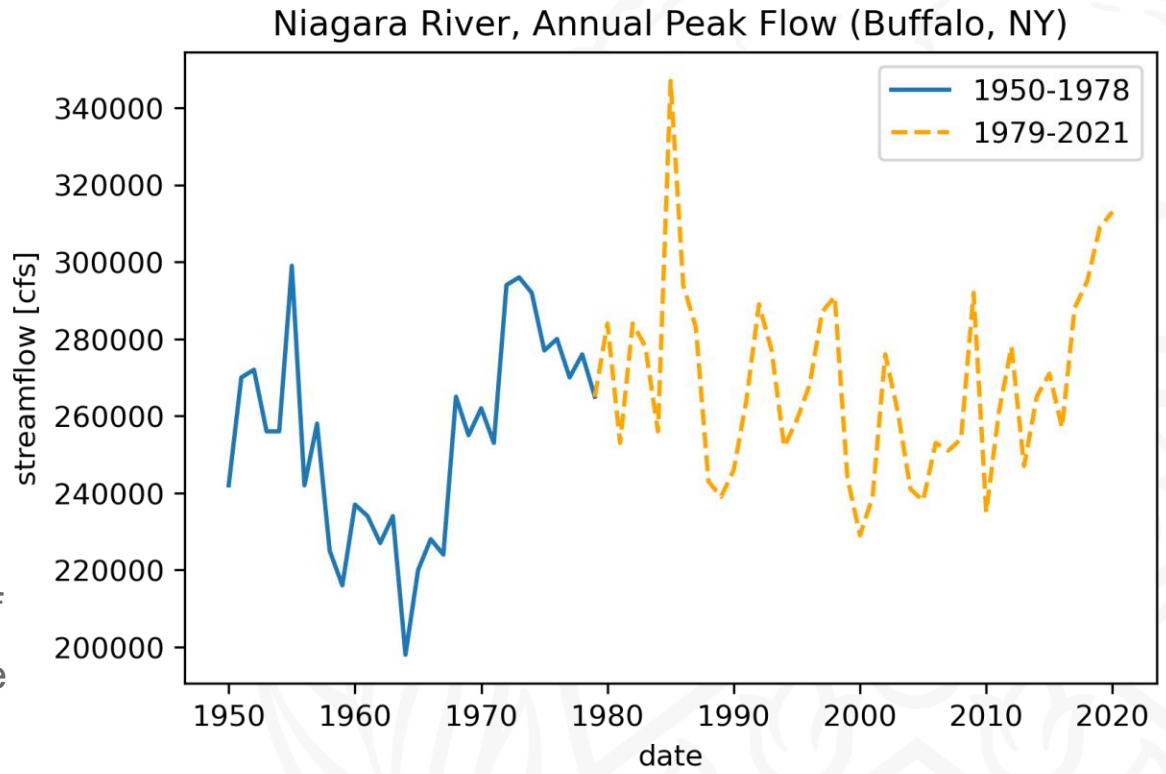
Open-Source Hydro Data Analytics

Sep 24th 2025

University at Buffalo The State University of New York

Hypothesis testing

- Why do we need hypothesis testing?
  - **Reproducible** - they ensure that every analysis of a dataset using the same methods will arrive at the same result because computations can be checked and agreed upon by others.
  - **Quantitative** - They present a quantitative measure of the strength of the evidence (the p-value), allowing the decision to reject a hypothesis to be augmented by the risk of an incorrect decision.



Hypothesis testing

Hypothesis

- A premise or claim that we want to test

Null hypothesis (H_0)

- Currently accepted value for a parameter

Alternative hypothesis (H_a)

- Also called the Research Hypothesis.
- Involves the claims to be tested

Example:

It is believed that a candy machine makes chocolate bars that are on average 5g. A worker claims that the machine after maintenance no longer makes 5g bars. Write **H_0** and **H_a**.

$$H_0: \mu = 5g$$
$$H_a: \mu \neq 5g$$

H_0 & H_a are mathematical opposites.

Practice – Write H_o and H_a

- A company has stated that their straw machine makes straws that are 4mm in diameter. A worker believes that the machine no longer makes straws of this size and samples 100 straws to perform a hypothesis test with 99% confidence. Write **H_0** and **H_a**.

H_0: $\mu = 4mm$
H_a: $\mu \neq 4mm$

- The school board claims that at least 60% of students bring a phone to school. A teacher believes this number is too high and randomly samples 25 students to test at a level of significance of 0.02. Write **H_0** and **H_a**.

H_0: $P \geq 0.60$
H_a: $P < 0.60$

Test statistic

- Test statistics are calculated from sample data, that are used to decide whether 1) reject H_0 or 2) fail to reject H_0

Example:

Sample 50 bars (we cannot open all of them)
- Get the average value of the mass of the bar
- Calculate the test statistic

Bar 1: 5.2g
Bar 2: 4.9g
Bar 3: 5.5g
…
…
…
Bar 49: 6.4g
Bar 50: 5.5g

Sample mean: 5.5g

Statistically significant

- Where do we draw a line to make the decision?

H_0: $\mu = 5g$
H_a: $\mu \neq 5g$

Based on the sample size (50) and sample mean (5.5), are we confident enough to say that the machine is broken?

Level of Confidence (C)

- How confident are we in our decision?
  - Usually, we see 95% or 99% confidence level in research articles.

Level of Significance ($\alpha = 1 - C$)

- LOC=95% is equivalent to $\alpha = 0.05$

Type I/II errors

	H_0	
	True	False
Reject H_0	Type I error	✔
Fail to Reject H_0	✔	Type II error

Let's say that the null hypothesis H_0 is: John's used car is safe to drive.
(a) Which statement represents a type I error?
(b) Which statement represents a type II error?

1) John thinks that his car may be safe when in fact, it is not safe.
2) John thinks that his car may be safe when, in fact, it is safe.
3) John thinks that his car may not be safe when, in fact, it is not safe.
4) John thinks that his car may not be safe when, in fact, it is safe.

Type I/II errors

	H_0	
	True	False
Reject H_0	Type I error (4)	✔
Fail to Reject H_0	✔	Type II error (1)

Let's say that the null hypothesis H_0 is: John's used car is safe to drive.
(a) Which statement represents a type I error?
(b) Which statement represents a type II error?

1) John thinks that his car may be safe when in fact, it is not safe.
2) John thinks that his car may be safe when, in fact, it is safe.
3) John thinks that his car may not be safe when, in fact, it is not safe.
4) John thinks that his car may not be safe when, in fact, it is safe.

Type I/II errors

	H_0	
	True	False
Reject H_0	Type I error Probability = α	✔ Probability=1-β
Fail to Reject H_0	✔ Probability = 1-α	Type II error Probability=β

α denotes *level of significance*
$1 - \beta$ denotes *statistical power*

Statistical power (1-β)

- It quantifies the likelihood of a significance test detecting an effect when there actually is one.
- Power is mainly influenced by 1) sample size, 2) effect size, and 3) significance level

Possible outcomes of this tests

- Reject Null Hypothesis H_0

- Fail to Reject Null Hypothesis H_0

We compared p-values against significance levels to decide whether we ***reject*** or ***fail to reject***.

p-value

- The p-value measures how surprising the results are if our starting assumption (the null hypothesis **H_0**) is correct.
  - If $p \leq \alpha$, we reject null hypothesis
  - If $p > \alpha$, we fail to reject null hypothesis

How do we calculate p-values?

z-test

We sampled 1000 candies, and the following are sample metrics

```
sample_candy_weight_mean = np.mean(samples)
sample_candy_weight_sd = np.std(samples)
print("The mean weight for 1000 candy sample is %0.2f"%(sample_candy_weight_mean))
print("The mean  for 1000 candy sample is %0.2f"%(sample_candy_weight_sd))
```

```
The mean weight for 1000 candy sample is 5.50
The mean  for 1000 candy sample is 0.20
```

Then we calculate **z-score**! Z-score denotes the distance from the sample mean to the population mean in units of the standard error

$$z = \frac{\bar{X} - \mu}{\sigma}$$

Can be estimated using sample standard deviation when sample size is big

Example:

It is believed that a candy machine makes chocolate bars that are on average 5g. A worker claims that the machine after maintenance no longer makes 5g bars. Write **H_0** and **H_a**.

H_0: $\mu = 5g$
H_a: $\mu \neq 5g$

z-test

We sampled 1000 candies, and the following are test statistics

```
sample_candy_weight_mean = np.mean(samples)
sample_candy_weight_sd = np.std(samples)
print("The mean weight for 1000 candy sample is %0.2f"%(sample_candy_weight_mean))
print("The mean  for 1000 candy sample is %0.2f"%(sample_candy_weight_sd))
```

```
The mean weight for 1000 candy sample is 5.50
The mean  for 1000 candy sample is 0.20
```

Then we calculate **z-score**! Z-score denotes the distance from the sample mean to the population mean in units of the standard error

$$z = \frac{\bar{X} - \mu}{\sigma}$$

```
mu_0 = 5.0
z_metric = (sample_candy_weight_mean - mu_0) / sample_candy_weight_sd
print("z metric is %s"%z_metric)
```

```
z metric is 2.5204529757998637
```

Example:

It is believed that a candy machine makes chocolate bars that are on average 5g. A worker claims that the machine after maintenance no longer makes 5g bars. Write H_0 and H_a.

$$H_0: \mu = 5g$$
$$H_a: \mu \neq 5g$$

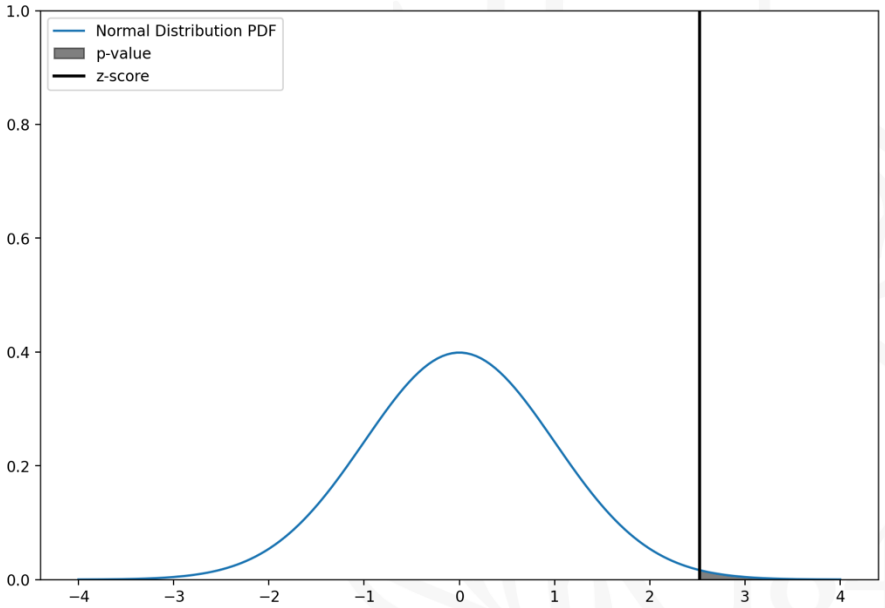
z-test

- We can simply calculate the p-value given a z-score

```
p_value = scipy.stats.norm.sf(z_metric)
p_value
```

```
0.005860194871506972 5
```

But what does p-value look like?



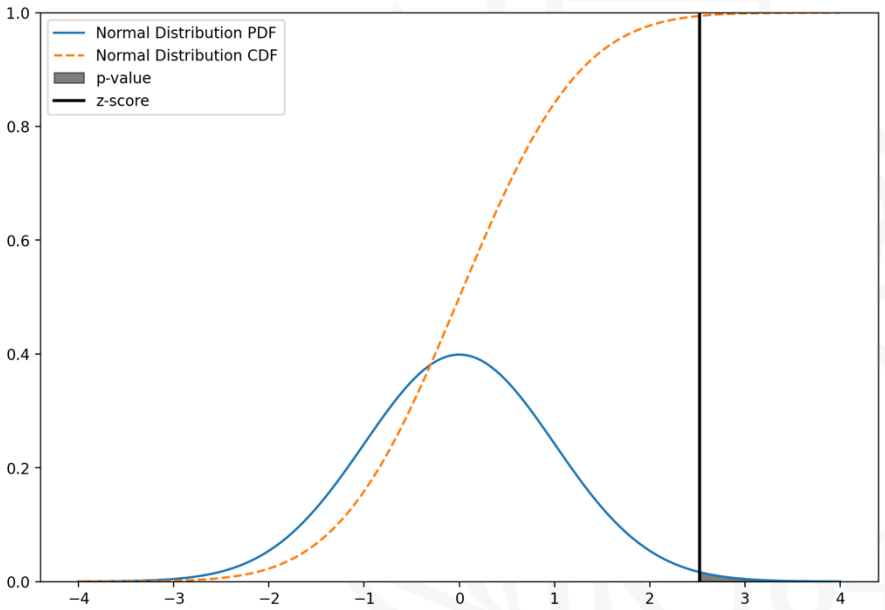
z-test

- We can simply calculate the p-value given a z-score

```
p_value = scipy.stats.norm.sf(z_metric)
p_value
```

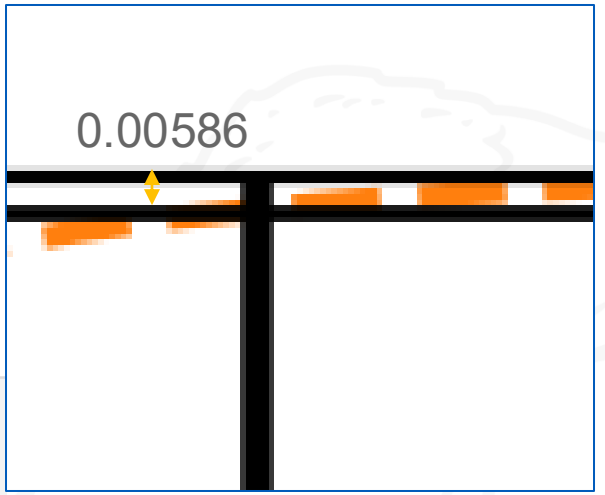
```
0.005860194871506972 5
```

But what does p-value look like?



z-test

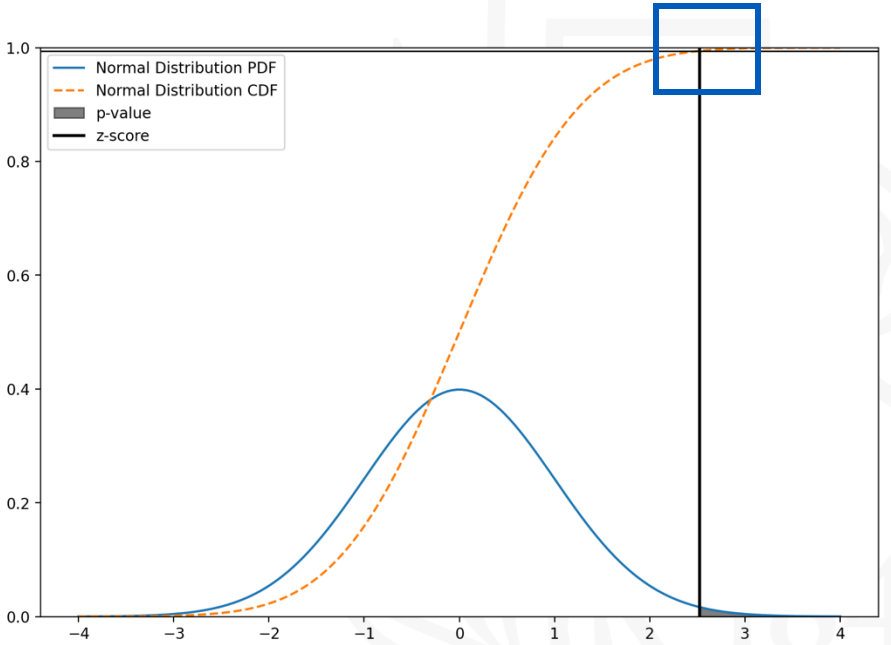
- We can simply calculate the p-value given a z-score




```
p_value = scipy.stats.norm.sf(z_metric)
p_value
```

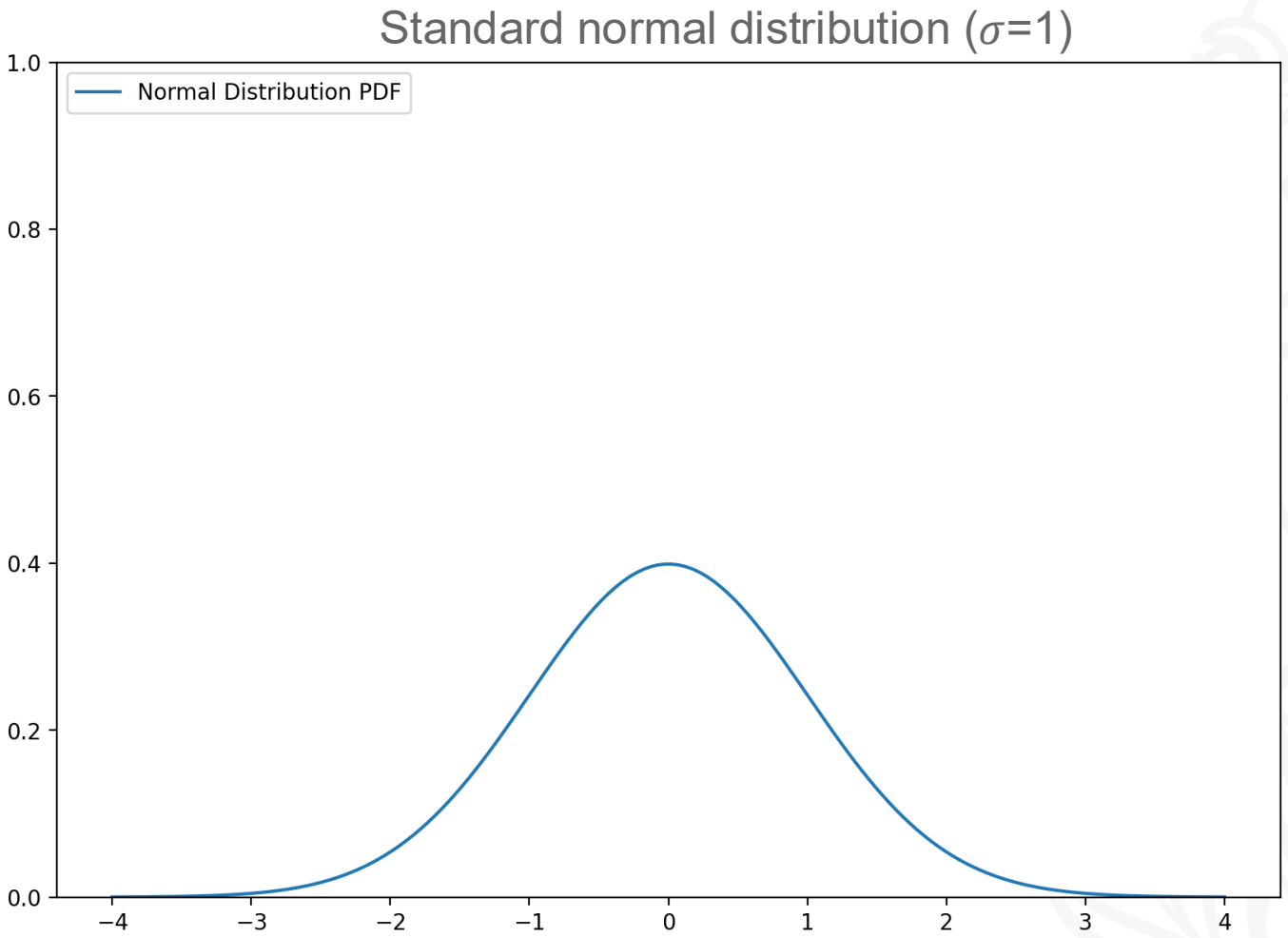
```
0.005860194871506972 5
```

But what does p-value look like?



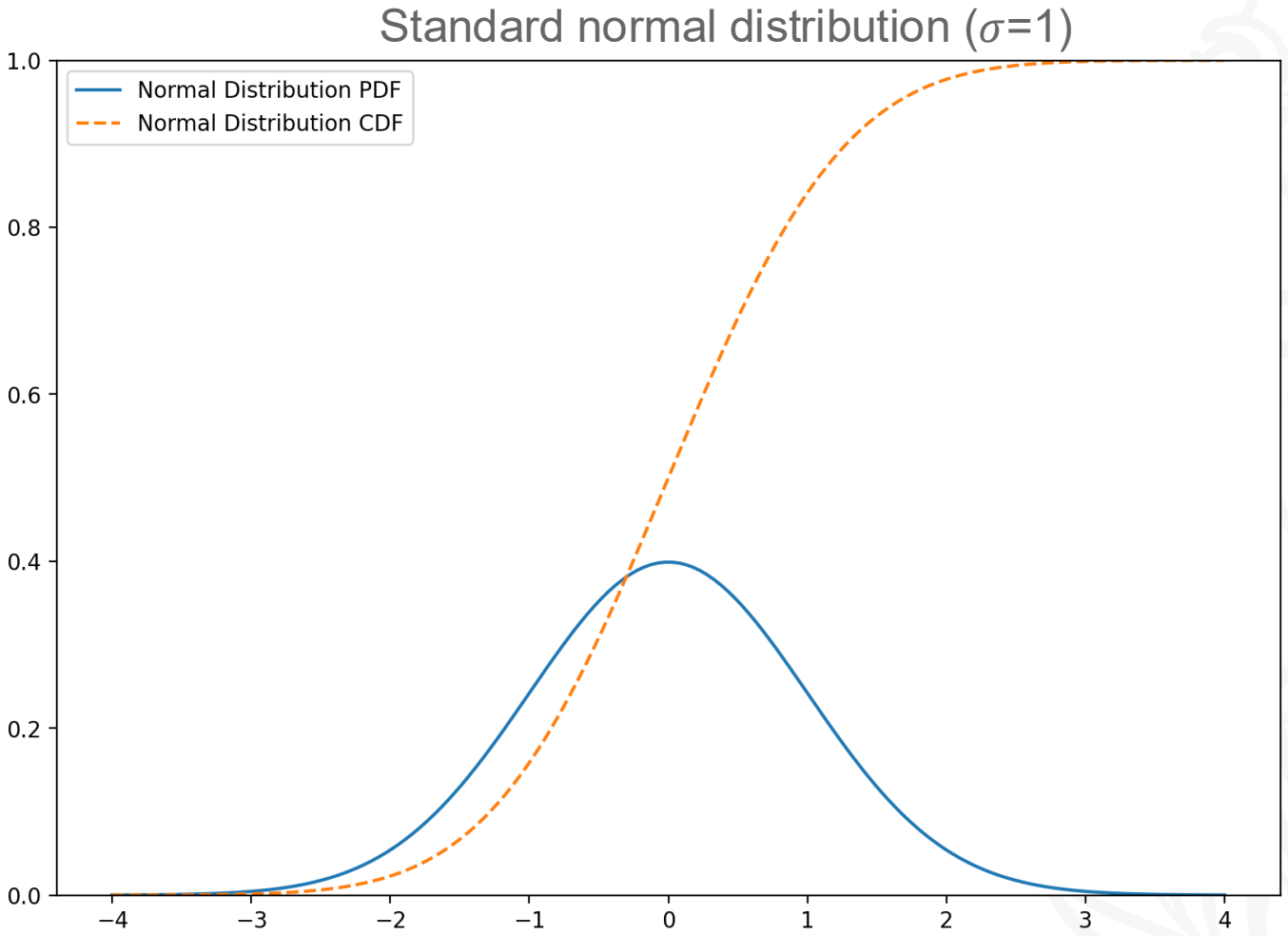
z-test visualization

What does the rejection region look like when α =0.05?



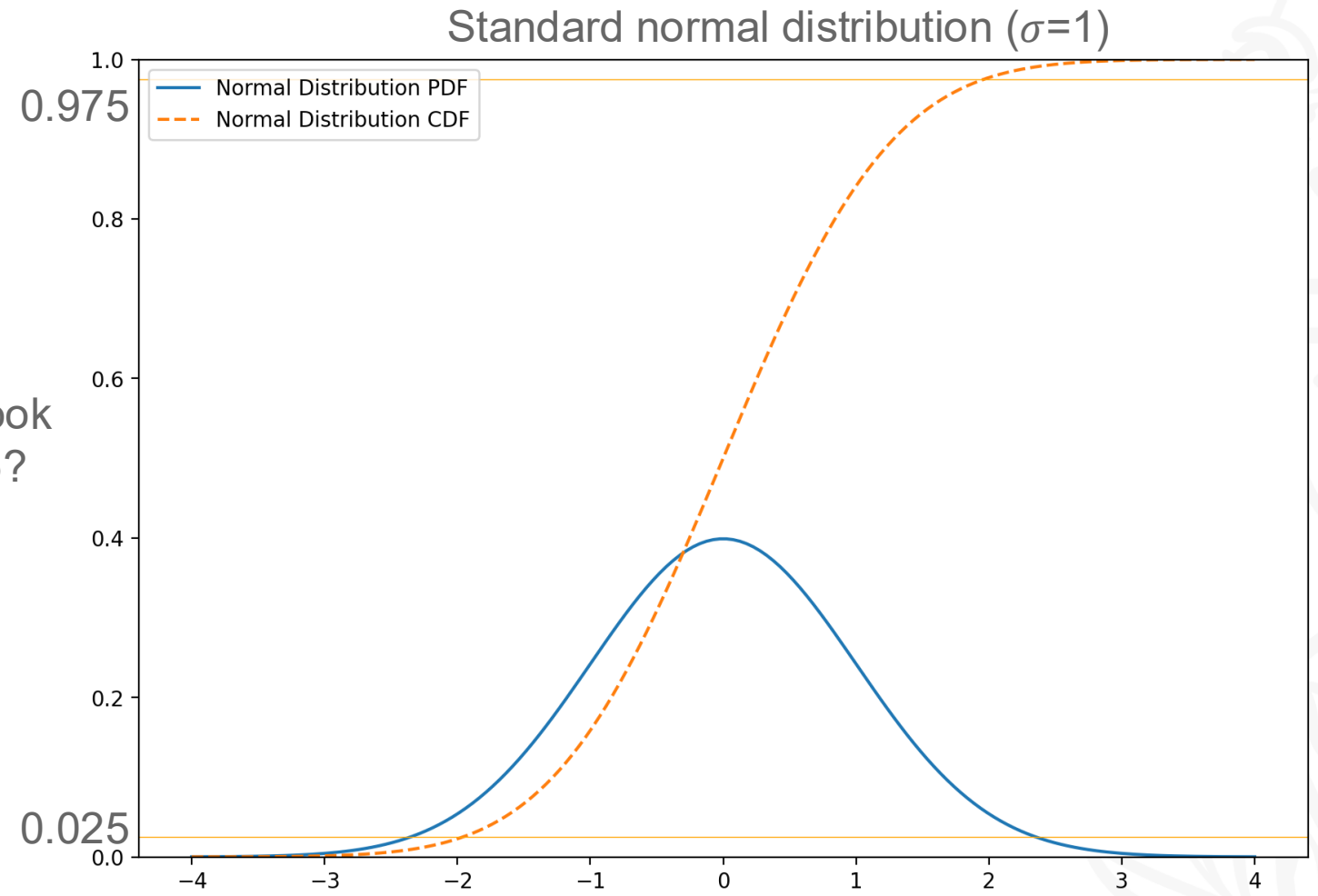
z-test visualization

What does the rejection region look like when α =0.05?



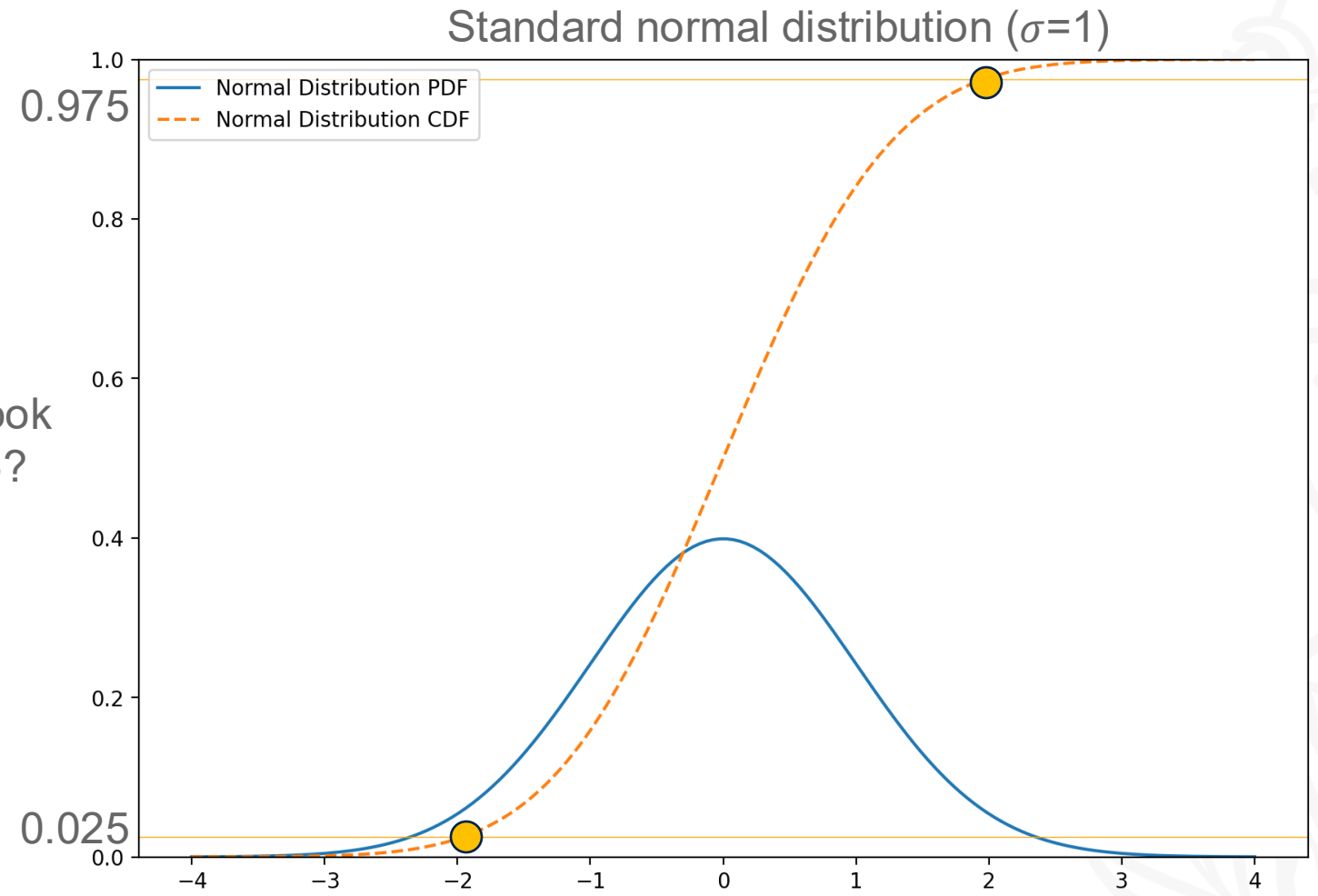
z-test visualization

What does the rejection region look like when α =0.05?



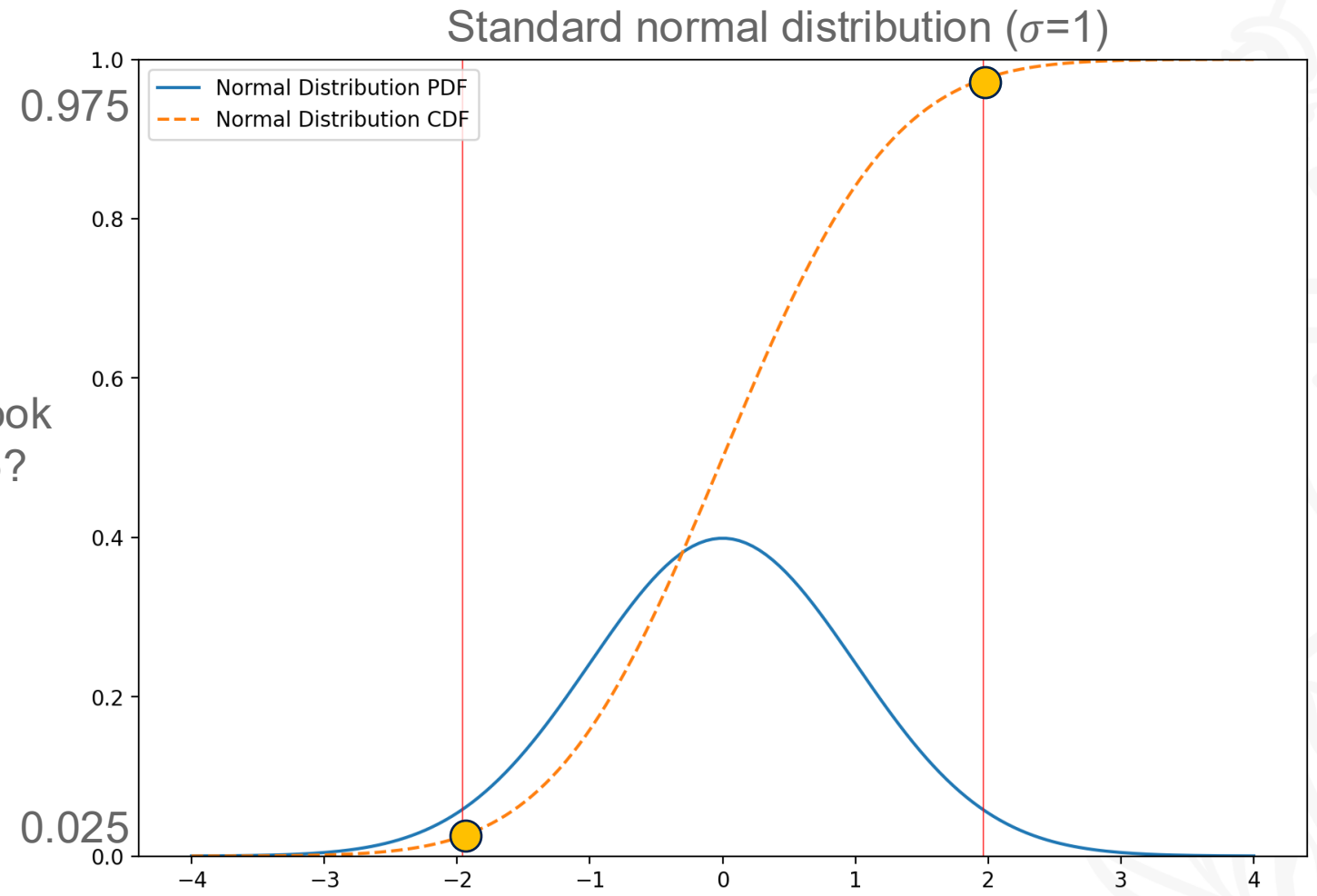
z-test visualization

What does the rejection region look like when α =0.05?



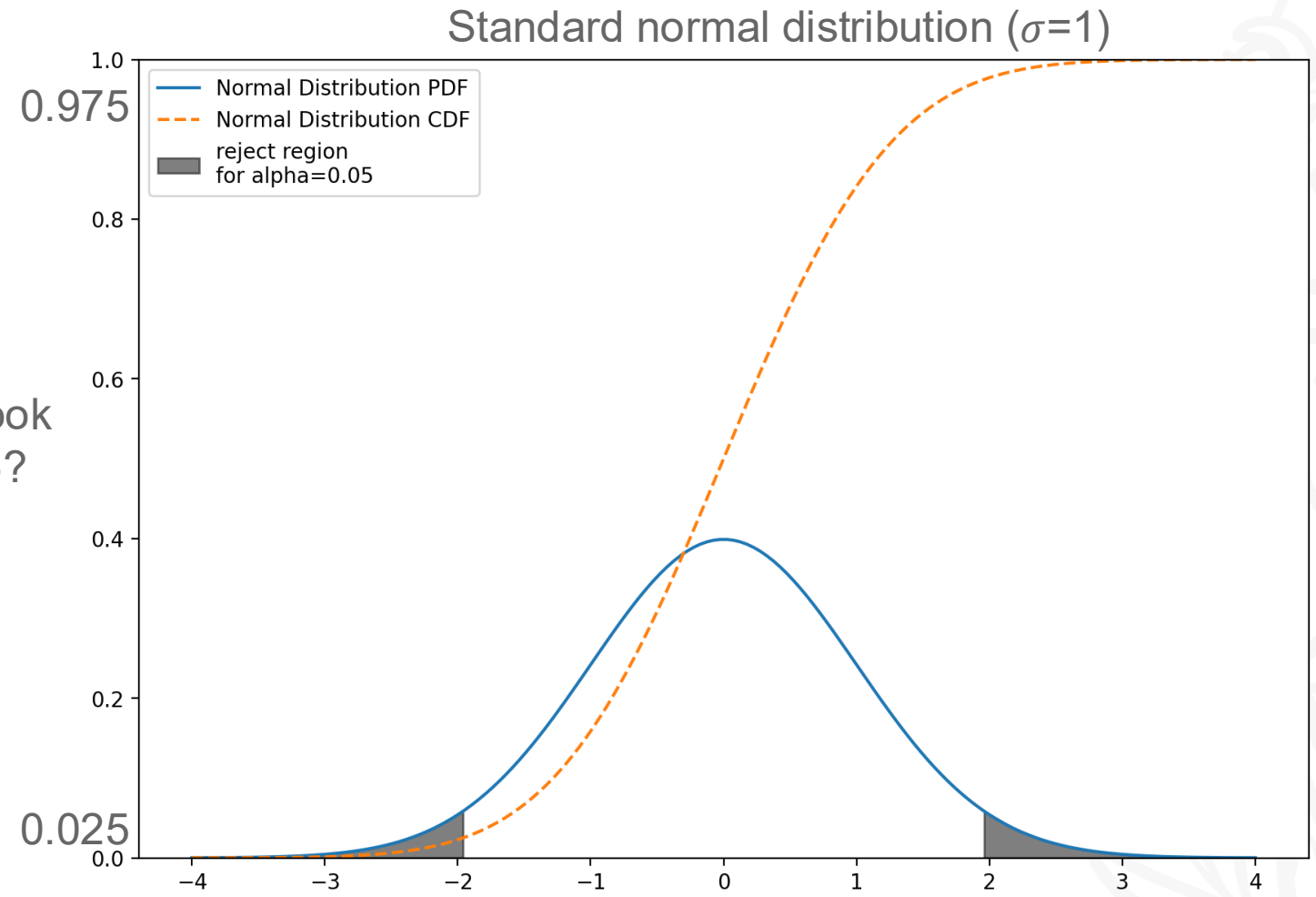
z-test visualization

What does the rejection region look like when α =0.05?



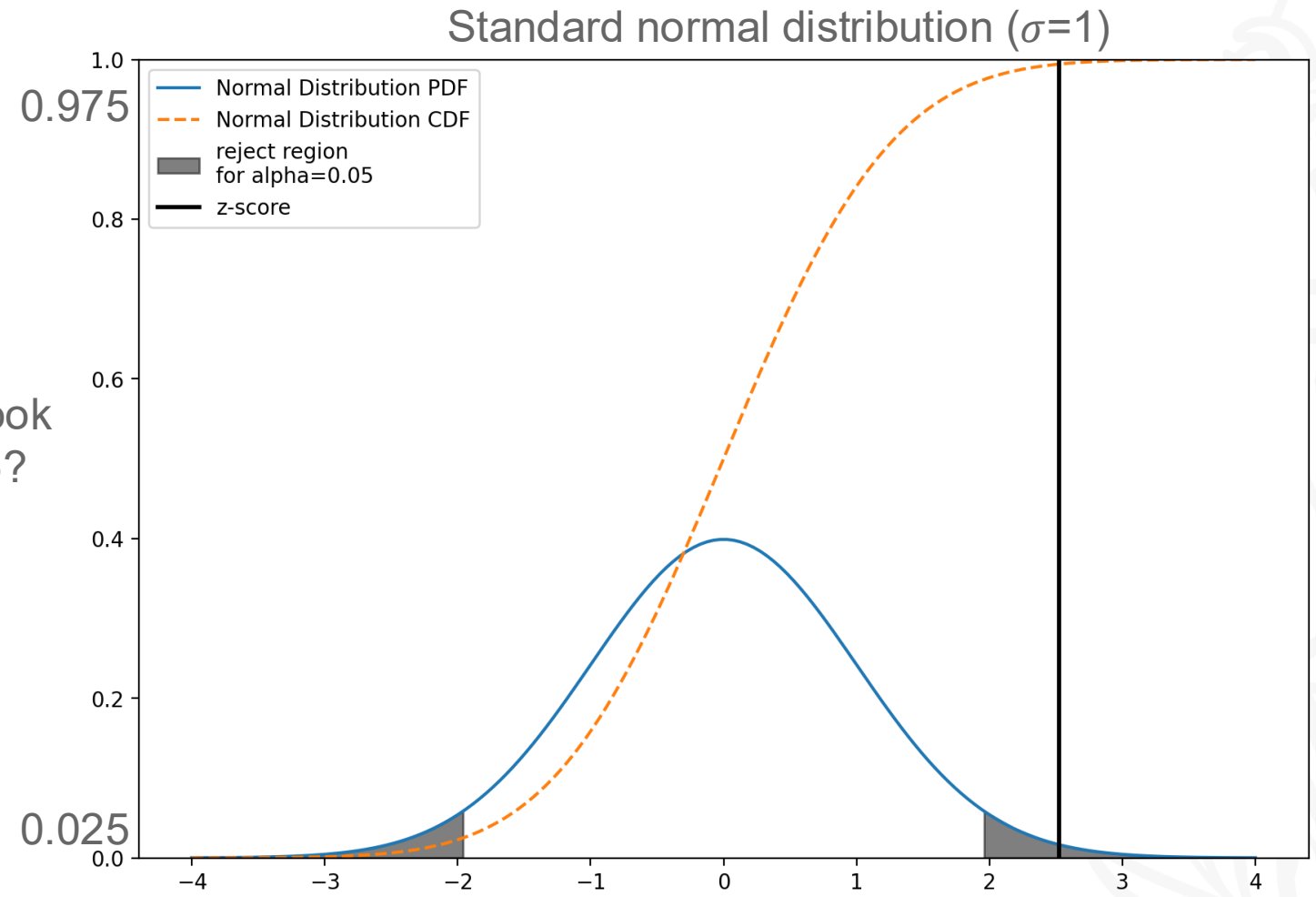
z-test visualization

What does the rejection region look like when α =0.05?



z-test visualization

What does the rejection region look like when α =0.05?



Z-score=2.52

Which falls within the rejection region

Thus, we can reject the null hypothesis!

Two sample T-test (Student's T-test)

1. Calculate sample mean

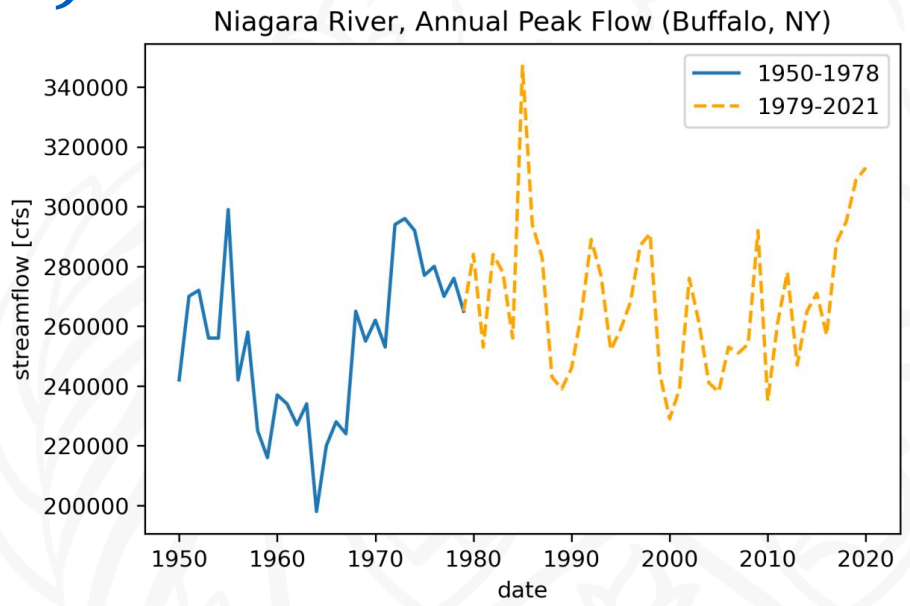
```
[149]: mean_peak_flow_pre_1978 = peak_flow_pre_1978['streamflow'].mean()
       mean_peak_flow_post_1978 = peak_flow_post_1978['streamflow'].mean()
       print("Mean peak flow before 1978 is %0.2f cfs"%mean_peak_flow_pre_1978)
       print("Mean peak flow after 1978 is %0.2f cfs"%mean_peak_flow_pre_1978)

       Mean peak flow before 1978 is 254100.00 cfs
       Mean peak flow after 1978 is 254100.00 cfs
```

2. Calculate sample standard deviation

```
[150]: sd_peak_flow_pre_1978 = peak_flow_pre_1978['streamflow'].std()
       sd_peak_flow_post_1978 = peak_flow_post_1978['streamflow'].std()
       print("SD for peak flow before 1978 is %0.2f cfs"%sd_peak_flow_pre_1978)
       print("SD for peak flow after 1978 is %0.2f cfs"%sd_peak_flow_post_1978)

       SD for peak flow before 1978 is 26227.58 cfs
       SD for peak flow after 1978 is 24428.09 cfs
```



$H_0: \overline{X_1} = \overline{X_2}$

$H_a: \overline{X_1} \neq \overline{X_2}$

We only have a limited number of samples!

Two sample T-test (Student's T-test)

3. Calculate t-score

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

s_p is the pooled standard deviation of the two samples

[165]:
```
n_pre_1978 = len(peak_flow_pre_1978)
n_post_1978 = len(peak_flow_post_1978)
print("n1 = %i"%n_pre_1978)
print("n2 = %i"%n_post_1978)
```

```
n1 = 30
n2 = 42
```

[166]:
```
sp = np.sqrt(((n_pre_1978-1)*sd_peak_flow_pre_1978 **2 +
              (n_post_1978-1)*sd_peak_flow_post_1978 **2)/
              (n_pre_1978+n_post_1978-2))
print("pooled standard deviation is %0.2f cfs"%sp)
```

```
pooled standard deviation is 25189.20 cfs
```

[167]:
```
t_score = (mean_peak_flow_post_1978 - mean_peak_flow_pre_1978)/\
          (sp * np.sqrt(1/n_pre_1978+1/n_post_1978))
print("t-score is %0.2f"%t_score)
```

```
t-score is 2.31
```

Two sample T-test (Student's T-test)

3. Calculate t-score

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

4. Find the t-score threshold in lookup table

$\alpha = 0.05$ (Two-tailed test)
degree of freedom = n1+n2-2

Two sample T-

3. Calculate t-score

$$t = \frac{\bar{X}_1 -}{s_p \cdot \sqrt{\frac{1}{n_1}}}$$

where

$$s_p = \sqrt{\frac{(n_1 -}{}}$$

4. Find the t-score thresh

$\alpha = 0.05$ (Two-tailed test)
degree of freedom = 70

t-score threshold = 1.995

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Two sample T-test (Student's T-test)

3. Calculate t-score

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$$

where

```
#Studnt, n=70, significance level: 0.05, 2-tail
#equivalent to Excel TINV(0.05,70)
print(stats.t.ppf(1-0.025, n_pre_1978+n_post_1978-2))
```

```
1.994437111771186
```

4. Find the t-score threshold in lookup table

You can easily get t-score threshold using python!

$\alpha = 0.05$ (Two-tailed test)
degree of freedom = 70

t-score threshold = 1.995

Two sample T-test (Student's T-test)

3. Calculate t-score

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}}$$

4. Find the t-score threshold in lookup table

$\alpha = 0.05$ (Two-tailed test)
degree of freedom = 70

t-score threshold = 1.995

t-score = 2.31 > 1.995

We can reject null hypothesis!

The annual peak flow after 1978 is ***significantly different*** than the annual peak flow before 1978.

More hypothesis tests?

- You can take a full course just to focus on this topic!

STA 504 Introduction to Applied Statistics II

3 Credits, Spring Semester

Advanced presentation of statistical methods for comparing populations and estimating and testing associations between variables. Topics: Point estimation, confidence intervals, hypothesis testing, ANOVA models for 1, 2 and k way classifications, multiple comparisons, chi-square test of homogeneity, Fisher's exact test, McNemar's test, measures of association, including odds ratio, relative risks, Mantel-Haenszel tests of association, and standardized rates, repeated measures ANOVA, simple regression and correlation. This course includes a one-hour computing lab and emphasizes hands-on applications to datasets from the health-related sciences.

More hypothesis tests?

Considerations for selecting which test to perform

1. **Data type:** The type of data you have, such as whether it's continuous, binary, or categorical
2. **Hypothesis:** Whether the hypothesis is one-tailed or two-tailed
3. **Sample size:** The size of your sample
4. **Research question:** The scientific question you're trying to answer
5. **Statistical assumptions:** Whether your data meets certain assumptions, such as independence of observations and homogeneity of variance
6. **Study design:** Whether the study design is paired or unpaired

Statistical Methods in Water Resources (PDF), Helsel, et al., 2020 https://doi.org/10.3133/tm4A3

6 Steps – hypothesis testing

- All hypothesis tests follow the same six steps, which are discussed in the following sections:
  1. Choose the appropriate test and review its assumptions.
  2. Establish the null and alternative hypotheses, H_0 and H_a.
  3. Decide on a significance level, α, or confidence level, C.
  4. Compute the test statistic from the data.
  5. Compute the p-value.
  6. Reject the null hypothesis if $p \leq \alpha$; fail to reject if $p > \alpha$.