Hypothesis Testing

Continued

Fun way to find critical Z

- use numpy to make data that is drawn from a proper gaussian distribution
- find the value of the 97.5 percentile!
- ^^^ is for alpha level of 0.05: 1 - 0.05/2 = 97.5

```
#%% Calculate Critcal Z

alpha = 0.05
# Make gaussian data
mu, sigma = 0, 1 # sample mean and SEM
normdata = np.random.normal(mu, sigma, 10000000)
print(np.quantile(normdata, 1 - alpha/2))
```

Hypothesis Testing

Saying with some degree of certainty that a baseline hypothesis is true or false

- example, the weight of 100 shampoo bottles are observed…
  - mean is 199.94g
  - SD is 0.613
    - recall SD is square root of variance;
- baseline, or null hypothesis is that the **sampling average weight** is 200g
- problem: determine if the observed mean would reasonably be observed if the true mean were 200g
  - let the confidence level be set to 95%
  - Critical Z for 95% percent confidence is ±1.96
  - Use the observed mean to compute the **standardized test statistic**
  - If T is within the bounds we choose, the baseline null hypothesis stands!

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Hypothesis Testing

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Saying with some degree of certainty that a baseline hypothesis is true or false

- example, the weight of 100 shampoo bottles are observed…
  - mean is 199.94g
  - SD is 0.613
    - recall SD is square root of variance;
- baseline, or null hypothesis is that the **sampling average weight** is 200g
- problem: determine if the observed mean would reasonably be observed if the true mean were 200g
  - let the confidence level be set to 95%
  - Critical Z for 95% percent confidence is ±1.96
  - Use the observed mean to compute the **standardized test statistic**
  - If T is within the bounds we choose, the baseline null hypothesis stands!

T = (199.94 - 200) / (0.613 / 10)

T = -0.979

CI = ± 1.96

T is within the CI, therefore the baseline hypothesis is not rejected!

The observed data do not disprove the hypothesis that 200g is the true weight of a shampoo bottle.

For observations > 30 !

- compared to 0

$$T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

$$T = \frac{\bar{x}_1 - \mu_1}{\sqrt{\frac{s_1^2}{n_1}}}$$

- compared to independent sample

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Google form

https://forms.gle/smEUuRtLiQnfxAJT6

Look at Receipt Data

What hypothesis might you explore?

Plotting basics

- use matplotlib → import matplotlib.pyplot as plt
- reserve the space → plt.subplots()
- designate plot type → plt.hist(data)
- add labels → plt.xlabel("x axis label")
- add labels → plt.ylabel("y axis label"
- add title → plt.title("title of plot")

```python
import matplotlib.pyplot as plt

plt.subplots() # start a new figure
plt.boxplot(normdata) # box plot
plt.xlabel("dataset") # label
plt.ylabel("value") # label
plt.title("Box plot") # title

plt.subplots() # start a new figure
plt.hist(normdata, bins = 100) # historgram
plt.xlabel("value") # label
plt.ylabel("frequency") # label
plt.title("Historgram") # title

plt.subplots() # start a new figure
plt.violinplot(normdata) # violin plot
plt.xlabel("dataset") # label
plt.ylabel("frequency") # label
plt.title("Violin Plot") # title

plt.subplots() # start a new figure
plt.plot(normdata) # line plot
plt.xlabel("order") # label
plt.ylabel("value") # label
plt.xlim((0, 100)) # limit the view
plt.title("Line plot") # title

counts, bins = np.histogram(normdata)
labels = ['1','2','3','4','5','6','7','8','9','10']
plt.subplots() # start a new figure
plt.pie(counts, labels = labels) # pie chart
plt.title("Pie Chart")
```