50.034 - Introduction to Probability and Statistics

Week 11 – Cohort Class

January–May Term, 2019

CC Week 11 Assigned Readings: 10.5

Outline of Cohort Class

- Exercise on p-value
- Example on Neyman–Pearson's lemma and most powerful test
- **Class Activity:** Simpson's Paradox

Recall: p-value

Let $T = T(X_1, \ldots, X_n)$ be a fixed statistic of a random sample $\{X_1, \ldots, X_n\}$ of observable R.V.'s with unknown parameter θ.

Let $\mathcal{H} = \{\mathcal{H}_c\}_{c \in \mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represents the hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$, test statistic T, and rejection region $[c, \infty)$. [Note: Ω_0 is a fixed subset.]

- Let α_c be the **size** of each $\mathcal{H}_c$, i.e. α_c is the smallest possible significance level for $\mathcal{H}_c$. (Different values of c give different sizes.)

Definition: Given some observed values $X_1 = x_1, \ldots, X_n = x_n$, let $t = T(x_1, \ldots, x_n)$ be the corresponding observed value of T. Then as c varies over $\mathbb{R}$, the smallest possible size α_c for which H_0 will be rejected given the observed value t, is called the p-value of $\mathcal{H}$.

- **Note:** The p-value depends on the observed values $x_1, \ldots, x_n$.
- **Interpretation:** If α is the p-value of $\mathcal{H}$, then it means that our experimental data is sufficient evidence to reject the null hypothesis H_0, whenever the value of c is chosen such that $\mathcal{H}_c$ has a significance level $\alpha_c \geq \alpha$.

Exercise 1 (20 mins)

Let $X_1, \ldots, X_{100}$ be iid normal observable R.V.'s with unknown mean μ and known variance 25.
Let $\mathcal{H} = \{\mathcal{H}_c\}_{c \in \mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represent a hypothesis test with null hypothesis $H_0 : \mu = 10$, test statistic $T = |\overline{X}_{100} - 10|$, and rejection region $[c, \infty)$.
Suppose that the observed values $X_1 = x_1, \ldots, X_{100} = x_{100}$ satisfy

$$\frac{x_1 + \cdots + x_{100}}{100} = 11.2.$$

What is the p-value of $\mathcal{H}$?

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{1/2}} \exp\left(-\tfrac{1}{2}u^2\right) du$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	0.60	0.7257	1.20	0.8849	1.80	0.9641	2.40	0.9918
0.01	0.5040	0.61	0.7291	1.21	0.8869	1.81	0.9649	2.41	0.9920
0.02	0.5080	0.62	0.7324	1.22	0.8888	1.82	0.9656	2.42	0.9922
0.03	0.5120	0.63	0.7357	1.23	0.8907	1.83	0.9664	2.43	0.9925
0.04	0.5160	0.64	0.7389	1.24	0.8925	1.84	0.9671	2.44	0.9927
0.05	0.5199	0.65	0.7422	1.25	0.8944	1.85	0.9678	2.45	0.9929
0.06	0.5239	0.66	0.7454	1.26	0.8962	1.86	0.9686	2.46	0.9931
0.07	0.5279	0.67	0.7486	1.27	0.8980	1.87	0.9693	2.47	0.9932
0.08	0.5319	0.68	0.7517	1.28	0.8997	1.88	0.9699	2.48	0.9934
0.09	0.5359	0.69	0.7549	1.29	0.9015	1.89	0.9706	2.49	0.9936
0.10	0.5398	0.70	0.7580	1.30	0.9032	1.90	0.9713	2.50	0.9938
0.11	0.5438	0.71	0.7611	1.31	0.9049	1.91	0.9719	2.52	0.9941

Exercise 1 - Solution

The given **null hypothesis** is $H_0 : \mu = 10$.

If indeed H_0 is true, then $\mathbf{E}[\overline{X}_{100}] = 10$ and $\mathrm{var}(\overline{X}_{100}) = \frac{25}{100} = 0.25$, so $\frac{\overline{X}_{100}-10}{\sqrt{0.25}} = 2(\overline{X}_{100} - 10)$ is a standard normal R.V.

We are told that H_0 is rejected if $|\overline{X}_{100} - 10| \geq c$. Note that

$$\begin{aligned}
\Pr(|\overline{X}_{100} - 10| \geq c) &= 1 - \Pr(-c \leq \overline{X}_{100} - 10 < c) \\
&= 1 - \Pr(-2c \leq 2(\overline{X}_{100} - 10) < 2c) \\
&= 1 - \Big(\Phi(2c) - \Phi(-2c)\Big) \\
&= 1 - \Big(\Phi(2c) - (1 - \Phi(2c))\Big) \\
&= 2 - 2 \cdot \Phi(2c),
\end{aligned}$$

where $\Phi(z)$ denotes the standard normal cdf.

Exercise 1 - Solution (continued)

Previous slide: $\Pr(|\overline{X}_{100} - 10| \geq c) = 2 - 2 \cdot \Phi(2c)$.

Let $\overline{x}_{100} = \frac{x_1 + \cdots + x_{100}}{100}$. Since we are given that $\overline{x}_{100} = 11.2$, it follows that H_0 will be rejected whenever $c \geq 1.2$.
Note that $c \geq 1.2$ corresponds to significance level

$$2 - 2 \cdot \Phi(c) \leq 2 - 2 \cdot \Phi(2.4) \approx 2 - 2(0.9918) = 0.0164.$$

Therefore, the p-value of $\mathcal{H}$ is 0.0164.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{1/2}} \exp\left(-\tfrac{1}{2}u^2\right) du$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	0.60	0.7257	1.20	0.8849	1.80	0.9641	2.40	0.9918
0.01	0.5040	0.61	0.7291	1.21	0.8869	1.81	0.9649	2.41	0.9920
0.02	0.5080	0.62	0.7324	1.22	0.8888	1.82	0.9656	2.42	0.9922
0.03	0.5120	0.63	0.7357	1.23	0.8907	1.83	0.9664	2.43	0.9925
0.04	0.5160	0.64	0.7389	1.24	0.8925	1.84	0.9671	2.44	0.9927
0.05	0.5199	0.65	0.7422	1.25	0.8944	1.85	0.9678	2.45	0.9929
0.06	0.5239	0.66	0.7454	1.26	0.8962	1.86	0.9686	2.46	0.9931
0.07	0.5279	0.67	0.7486	1.27	0.8980	1.87	0.9693	2.47	0.9932
0.08	0.5319	0.68	0.7517	1.28	0.8997	1.88	0.9699	2.48	0.9934
0.09	0.5359	0.69	0.7549	1.29	0.9015	1.89	0.9706	2.49	0.9936
0.10	0.5398	0.70	0.7580	1.30	0.9032	1.90	0.9713	2.50	0.9938
0.11	0.5438	0.71	0.7611	1.31	0.9049	1.91	0.9719	2.52	0.9941
0.12	0.5478	0.72	0.7642	1.32	0.9066	1.92	0.9726	2.54	0.9945
0.13	0.5517	0.73	0.7673	1.33	0.9082	1.93	0.9732	2.56	0.9948
0.14	0.5557	0.74	0.7704	1.34	0.9099	1.94	0.9738	2.58	0.9951
0.15	0.5596	0.75	0.7734	1.35	0.9115	1.95	0.9744	2.60	0.9953

Exercise 2 (solution provided)

Let $\{X_1, \ldots, X_{20}\}$ be a random sample of observable exponential random variables with unknown parameter θ.

Suppose we are given that the parameter space of θ contains only two possible values 1 and 2.

Find a most powerful hypothesis test with significance level 0.05, such that its null hypothesis is $H_0 : \theta = 1$. Please give a complete description of this most powerful hypothesis test $\mathcal{H}$, including the test statistic and rejection region of $\mathcal{H}$.

Hint: If Z is a gamma R.V. with parameters 20 and 1, then $\Pr(Z < 13.25465) = 0.05$.

The purpose of this exercise is to understand the solution.

Exercise 2 - Solution

We are given the following hypotheses:

$$H_0 : \theta = 1 \quad \text{(null hypothesis)};$$
$$H_1 : \theta = 2 \quad \text{(alternative hypothesis)};$$

Since both the null hypothesis and the alternative hypothesis are simple, we can apply Neyman–Pearson's lemma.

- Neyman–Pearson's lemma says that if we want a most powerful hypothesis test with significance level 0.05, then we can let the hypothesis test $\mathcal{H}$ be based on the likelihood ratio.

Given a vector $\mathbf{x} = (x_1, \ldots, x_{20})$ of observed values, suppose we denote the likelihood function of θ (given $\mathbf{x}$) by $\mathcal{L}(\theta|\mathbf{x})$. Then the likelihood ratio of $\mathbf{x}$ is

$$\frac{\mathcal{L}(2|\mathbf{x})}{\mathcal{L}(1|\mathbf{x})}.$$

- The alternative hypothesis corresponds to the numerator, and the null hypothesis corresponds to the denominator.

Exercise 2 - Solution (continued)

Next, we shall determine the likelihood ratio. Remember, we are given that each X_i is exponential with parameter θ.
The conditional marginal pdf of each X_i given $\theta = 2$ is

$$f_{X_i}(x_i|\theta = 2) = \begin{cases} 2e^{-2x_i}, & \text{if } x_i \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Thus the likelihood function $\mathcal{L}(\theta|\mathbf{x})$ at $\theta = 2$ is

$$\begin{aligned} \mathcal{L}(2|\mathbf{x}) &= f_{20}(\mathbf{x}|\theta = 2) = f_{X_1}(x_1|\theta = 2) \cdots f_{X_{20}}(x_{20}|\theta = 2) \\ &= \begin{cases} 2^{20}e^{-2(x_1+\cdots+x_{20})}, & \text{if } x_i \geq 0 \text{ for all } i; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

- **Note:** The joint conditional pdf $f_{20}(\mathbf{x}|\theta = 2)$ is a product of the marginal conditional pdf's $f_{X_i}(x_i|\theta = 2)$, since $X_1, \ldots, X_{20}$ are iid.

Exercise 2 - Solution (continued)

Similarly, the conditional marginal pdf of each X_i given $\theta = 1$ is

$$f_{X_i}(x_i|\theta = 1) = \begin{cases} e^{-x_i}, & \text{if } x_i \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Thus the likelihood function $\mathcal{L}(\theta|\mathbf{x})$ at $\theta = 1$ is

$$\begin{aligned} \mathcal{L}(1|\mathbf{x}) &= f_{20}(\mathbf{x}|\theta = 1) = f_{X_1}(x_1|\theta = 1) \cdots f_{X_{20}}(x_{20}|\theta = 1) \\ &= \begin{cases} e^{-(x_1+\cdots+x_{20})}, & \text{if } x_i \geq 0 \text{ for all } i; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently, the likelihood ratio is

$$\frac{\mathcal{L}(2|\mathbf{x})}{\mathcal{L}(1|\mathbf{x})} = \begin{cases} 2^{20}e^{-(x_1+\cdots+x_{20})}, & \text{if } x_i \geq 0 \text{ for all } i; \\ 0, & \text{otherwise.} \end{cases}$$

Thus, by **Neyman–Pearson's lemma**, we can use the test statistic

$$\Lambda(X_1, \ldots, X_{20}) = 2^{20}e^{-(X_1+\cdots+X_{20})}.$$

Exercise 2 - Solution (continued)

What we know so far: Neyman–Pearson's lemma tells us that if we want a most powerful test $\mathcal{H}$ with the given hypotheses H_0 and H_1, then we can use the test statistic

$$\Lambda(X_1, \ldots, X_{20}) = 2^{20} e^{-(X_1 + \cdots + X_{20})}.$$

and some rejection region of the form (k, ∞) for some constant k. We want $\mathcal{H}$ to be a most powerful test at significance level 0.05, which means that

$$\Pr(\Lambda > k | \theta = 1) = \Pr(2^{20} e^{-(X_1 + \cdots + X_{20})} > k | \theta = 1) = 0.05.$$

Let $Y = X_1 + \cdots + X_{20}$, and note that

$$\begin{aligned}
\Pr(2^{20} e^{-Y} > k | \theta = 1) &= \Pr(\ln(2^{20} e^{-Y}) > \ln k | \theta = 1) \\
&= \Pr\left((20 \ln 2 - Y) > \ln k | \theta = 1\right) \\
&= \Pr\left(Y < (20 \ln 2 - \ln k) | \theta = 1\right).
\end{aligned}$$

Exercise 2 - Solution (continued)

Given that $\theta = 1$, it means each X_i is exponential with parameter 1.

Question: What is the distribution of Y?

- **Useful observation:** An exponential R.V. with parameter θ is precisely a gamma R.V. with parameters 1 and θ (Lecture 15).
- **Useful Result:** (First appeared as hint in Homework 7)
  If $Y_1, \ldots, Y_k$ are independent R.V.'s, where each Y_i is gamma with parameters α_i and β, then the sum $Y_1 + \cdots + Y_k$ is gamma with parameters $\alpha_1 + \cdots + \alpha_k$ and β.

Consequence: Given $\theta = 1$, then $Y = X_1 + \cdots + X_{20}$ is a gamma R.V. with parameters 20 and 1.

Thus, from the hint,

$$\Pr(\Lambda > k|\theta = 1) = \Pr\Big(Y < (20 \ln 2 - \ln k)|\theta = 1\Big) = 0.05$$

So $20 \ln 2 - \ln k \approx 13.25465$, which gives $k \approx 1.8340$.

Exercise 2 - Solution (continued)

Therefore, we found a most powerful test $\mathcal{H}$ at significance level 0.05 with null hypothesis $H_0 : \theta = 1$ and alternative hypothesis $H_1 : \theta = 2$. The test statistic of $\mathcal{H}$ is

$$\Lambda(X_1, \ldots, X_{20}) = 2^{20} e^{-(X_1 + \cdots + X_{20})},$$

and the rejection region is the interval $(1.8340, \infty)$.

Class Activity: Paper ball toss (30 mins)

Student (Contestant A) versus Professor (Contestant B)

Goal: Score as many points as possible.

- Try your best to throw a paper ball into a bin.
- Each successful throw is worth 1 point.

Two Levels of Difficulty:

- Easy: You stand 5 cm away from the bin.
- Hard: You stand 5 m away from the bin.

Important Game Rule:

- Contestant A: n_A hard attempts, m_A easy attempts.
- Contestant B: n_B hard attempts, m_B easy attempts.
- The professor can decide the values of n_A, m_A, n_B, m_B.

Is it possible for Contestant A to have a higher shooting percentage than Contestant B for each level of difficulty, but still have a lower overall shooting percentage (both easy and hard levels combined)?

Discussion: What is Simpson's Paradox?

Key Idea: A trend can seemingly reverse after aggregation.

- The student can have a higher shooting percentage for each of the two difficulty levels, but can still have a lower overall shooting percentage when the attempts from both levels are aggregated.
- Given several datasets, each of which shows an "obvious trend" (e.g. A is better than B), it is sometimes possible to aggregrate these datasets into one larger dataset, where the "seemingly obvious" trend is reversed (e.g. B is better than A) on the larger combined dataset.
- Conversely, given a dataset that shows an "obvious" trend (e.g. A is better than B), it is sometimes possible to split the dataset into many small subsets, such that the "seemingly obvious" trend is reversed (e.g. B is better than A) on each small subset.

Note: Simpson's Paradox is not actually a paradox!

- It is just surprising to someone who has not seen it before.

UC Berkeley Gender Bias Example

One of the most famous real-world examples of Simpson's paradox:

- Gender bias in Berkeley's graduate school admissions in 1973.

	Men		**Women**	
	Applicants	**Admitted**	**Applicants**	**Admitted**
Total	8442	**42%**	4321	35%

- This suggests that men were more likely than women to be admitted!

Note: Every department conducts its own admissions exercise.

- When examining data from individual departments, it appears that 6 departments (out of 85) were biased against men, while 4 departments were biased against women.
- Further analysis shows that in fact there was a "small but statistically significant bias in favour of women".

UC Berkeley Gender Bias Example

Data from six largest departments in Berkeley:

Department	Men		Women	
	Applicants	**Admitted**	**Applicants**	**Admitted**
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

A research paper by *Bickel et al.* concluded from the data that:

> "Women tended to apply to competitive departments with low rates of admission even among qualified applicants (such as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry)"

Kidney Stone Treatment Example

Real-life medical study on two treatments for kidney stones:

- Categories for kidney stones: Small, Large.

	Treatment A	**Treatment B**
Small stones	**93% (81 of 87)**	87% (234 of 270)
Large stones	**73% (192 of 263)**	69% (55 of 80)
Both	78% (273 of 350)	**83% (289 of 350)**

- Treatment A: involves invasive surgical procedure
- Treatment B: involves only a small puncture
- It is believed that Treatment A is better than Treatment B.

1. Doctors tend to give Treatment A to more severe cases (large stones), and give Treatment B to mild cases (small stones).
2. Success rate is more strongly influenced by the severity of the case, rather than the choice of treatment.

Outcome: When Treatment B is given more frequently to less severe cases, it can appear to be more effective.

Arguing over Top Choices Example

This is a "made-up" example that could happen to you.

- You and your partner want to find the best resort for a holiday.
  - Your top choice: Seneca Resort.
  - Your partner's top choice: Cayuga Resort.
- From online reviews, you found the recommendation ratings:

	Seneca	**Cayuga**
Men	**50% (180 of 360)**	30% (50 of 150)
Women	**90% (36 of 40)**	80% (200 of 250)
Combined	54% (216 of 400)	**62.5% (250 of 400)**

- Your conclusion: Seneca is better than Cayuga.
- But your partner, using the same data, found that Cayuga is recommended by a higher percentage of all users.
- Your partner's conclusion: Cayuga is better than Seneca.

What should you do if both you and your partner are correct?

Hair Length Example (10 mins)

Let's consider the hair lengths of everyone in class.
Think of what happens as the population changes over time.

- Let ℓ_{ladies} be the average hair length of all ladies in this class.
- Let ℓ_{guys} be the average hair length of all guys in this class.
- Let ℓ_{overall} be the overall average hair length in this class.

Assume that $\ell_{\text{guys}} < \ell_{\text{overall}} < \ell_{\text{ladies}}$.

Now, suppose a guy walks into the classroom.

- His hair length is greater than ℓ_{guys}, but less than ℓ_{overall}.

Question: What happens to the new averages ℓ'_{overall}, ℓ'_{ladies}, ℓ'_{guys}?

- Even though $\ell'_{\text{ladies}} = \ell_{\text{ladies}}$ (no change) and $\ell'_{\text{guys}} > \ell_{\text{guys}}$ (increase), the overall trend is $\ell'_{\text{overall}} < \ell_{\text{overall}}$ (decrease).

Discussion Question: Suppose a group (5 guys, 1 lady) walk into the classroom. Is it possible that:

$$\ell''_{\text{ladies}} > \ell'_{\text{ladies}} \quad \text{and} \quad \ell''_{\text{guys}} > \ell'_{\text{guys}}, \quad \text{but} \quad \ell''_{\text{overall}} < \ell'_{\text{overall}}?$$

Example from Course Textbook

Example with equal number of men and women

A new treatment is compared versus a standard treatment.

All patients	Improved	Not Improved	% Improved
New Treatment	20	20	50%
Standard Treatment	24	16	**60%**

Male patients	Improved	Not Improved	% Improved
New Treatment	12	18	**40%**
Standard Treatment	3	7	30%

Female patients	Improved	Not Improved	% Improved
New Treatment	8	2	**80%**
Standard Treatment	21	9	70%

Note: There are 40 male patients and 40 female patients.

Discussion Question: What is happening here?

- **Hint 1:** Female patients tend to improve more than male patients, whichever treatment they get.
- **Hint 2:** Who gets the standard treatment more frequently?

Summary

- Exercise on p-value
- Example on Neyman–Pearson's lemma and most powerful test
- **Class Activity:** Simpson's Paradox

Reminders:

There is **mini-quiz 4** (15mins) next week during Cohort Class.

- Final mini-quiz! Tested on all materials from Lectures 15–20 and Cohort classes weeks 9–11. This is Week 11.

Make-up class for **next week**'s Friday's Cohort Class

- Originally on 19th April (Good Friday).
- Make-up: On 17th April (Wednesday), 2–4pm, CC14 (2.507).
  - So your mini-quiz 4 will be on Wednesday!
- Next Thursday's cohort classes are on as usual.