Experimental design and hypothesis testing

DSE 210

Outline

1. Design of experiments
   - Controlled experiments
   - Observational studies
2. Statistical hypothesis tests
   - The z statistic
   - The χ^2 statistic

Most of the examples I'll cover are from the textbook *Statistics* by David Freedman, Robert Pisani and Roger Purves.

A vaccine against polio

Timeline:

- 1916: First polio epidemic hit the US
- Over the next 40 years: hundreds of thousands of fatalities, especially children
- By the 1950s: several vaccines against polio were proposed
- 1954: Public Health Service and National Foundation for Infantile Paralysis (NFIP) were ready for real-world testing of a vaccine developed by Jonas Salk

How could this testing be done?

Salk vaccine: experimental design

Question: How about giving the vaccine to large numbers of children in 1954, and seeing if this led to a sharp drop in polio cases?

Bad idea: The incidence of polio varied from year to year. For instance, there were only half as many cases in 1953 than in 1952.

Controlled experiment:

- Need to deliberately leave some children unvaccinated: **controls**.
- Compare outcomes in the **treatment group** and the **control group**.

The NFIP experimental design:

- Chose two million children in selected school districts with high risk of polio, from the age groups most vulnerable (grades 1,2,3).
- Idea: would choose a million to vaccinate, and leave the rest unvaccinated, as controls.

The NFIP experimental design

How to partition the subjects into treatment and control groups?

- NFIP split it by grade level: grade 2 would get the vaccine, grades 1 and 3 would be controls.
- This is problematic. What if the incidence were higher in one grade than another? Such factors would **confound** the effect of treatment. Better idea: divide randomly.

A significant complication: parental consent.

- Those chosen for vaccination needed parental consent. Half the parents refused.
- Higher-income parents more likely to consent to treatment. Does this bias the study for or against the vaccine?
  Against. Children in less hygenic surroundings tend to contract mild cases while still protected by mother's antibodies, and this protects them later.

A better design

Textbook design: **randomized controlled double-blind** experiment.

- Control group needs to be from the same population as the treatment group.
  Therefore, select both from children whose parents consented to treatment.
- Choose the two groups at random from the same population.
  This is a **randomized controlled** experiment.
- Subjects should not know which group they are in.
  Therefore, children in the control group should be given a placebo.

Both designs were used: some school districts used the NFIP design, others used the double-blind design.

Salk vaccine: the results

For the double-blind randomized controlled experiment:

	Size	Rate (per 100K)
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

(The NFIP experiment showed a significantly weaker effect.)

How can we assess the significance of these numbers?

Historical controls

Sometimes, experiments compare outcomes for people receiving a new treatment to outcomes observed in the past without that treatment (historical controls). This is inferior to a randomized controlled design.

Example: What is the value of coronary bypass surgery for patients with coronary artery disease? Two studies, one with randomized controls and one with historical controls, reported these three-year survival rates:

	Randomized	Historical
Surgery	87.6%	90.0%
Controls	83.2%	71.1%

How might this discrepancy be explained?

Outline

1. Design of experiments
   - Controlled experiments
   - Observational studies
2. Statistical hypothesis tests
   - The z statistic
   - The χ^2 statistic

Observational studies

Two kinds of study:

- **Controlled experiment**: investigators decide who is in the treatment group and who is in the control group.
- **Observational study**: the subjects assign themselves to these two groups. The investigators just watch.

Example: studies on smoking are necessarily observational.

- Heart attacks, lung cancer, and various other diseases are more common among smokers than non-smokers.
- But perhaps there are other explanations: confounding factors that make people smoke and also make them sick.
- For instance: sex. Men are more likely to smoke than women, and are more likely to get heart disease.
- Or age: older people have different smoking habits and are more at risk for these diseases.

Careful observational studies have controlled for many confounding factors and together make a case that smoking does cause these diseases.

Cervical cancer and circumcision

For many years, cervical cancer was one of the most common cancers among women.

- Investigators looking for causes found that cervical cancer seemed to be rare among Jews.
- They also found it to be quite rare among Muslims.
- In the 1950s, various investigators concluded that circumcision of males protected against this cancer.

More recent studies suggest that cervical cancer is caused by human papilloma virus, which is sexually transmitted. More sexually active women, with more partners, are more likely to be exposed to it.

Ultrasound and low birthweight

Experiments on lab animals showed that ultrasound can cause low birthweight. Is this true for humans?

- Investigators at Johns Hopkins ran an observational study.
- They tried to adjust for various confounding factors.
- Even controlling for these, babies exposed to ultrasound on average had lower birthweight than those not exposed.

At that time, ultrasounds were used mostly during problem pregnancies: the common cause of the ultrasound and low birthweight. A later randomized controlled experiment showed no harm.

Statistical hypothesis testing

1. The z statistic
   - Testing the mean of a distribution
   - Testing whether two distributions have the same mean
2. The χ^2 statistic
   - Testing whether a sequence of $\{1, 2, \ldots, k\}$ outcomes comes from a particular k-sided die
   - Testing the independence of two variables

Example: new tax code

A senator introduces a change to the tax code that he claims is revenue-neutral. How can this be verified?

- See how this change would affect last year's tax returns.
- Pick 100 returns at random, look at the change in revenue of each.
- The average change is $-219.
- The standard deviation is $725.

Analyze this in the framework of **hypothesis testing**.

- **Null hypothesis**: The average change is $0.
- **Alternative hypothesis:** The average change is negative.

In order to discredit the null hypothesis, *argue by contradiction.*

- Assume the null is true.
- Compute a **statistic** that measures the difference between what is observed and what would be expected under the null.
- What is the chance of obtaining a statistic this extreme?

The z statistic

Pick 100 tax returns at random.

- The average change in revenue is $X = -219$ dollars.
- The standard deviation is $725.

How likely is X under the null?

- Recall null hypothesis: expected change is $0.
- Under the null, X would be normally distributed with mean 0 and standard deviation $725/10 = 72.5$.
- The observed X is ≈ 3 standard deviations from the mean: unlikely.

The z-**statistic** measures how many standard deviations away the observed value is from its expectation.

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{-219 - 0}{72.5} \approx -3$$

The probability of observing this under the null is the p-**value**.

This p-value is less than $1/1000$: strong evidence against the null.

Hypothesis testing: recap

The null hypothesis is what we are trying to discredit.

We do this by contradiction:

- Let the observation be denoted X.
- What is the distribution of X under the null?

If we would expect X to be normally distributed, we can use the z-statistic:

$$z = \frac{\text{observed } X - \text{expected } X}{\text{standard deviation of } X}$$

The p-value is the probability of seeing a value (at least) this extreme under the null. A small p-value is evidence against the null.

Example: an ESP demonstration

Charles Tart's experiments at UC Davis using the "Aquarius":

- Aquarius has an electronic random number generator
- Chooses one of four targets but doesn't reveal which
- The subject guesses which, and a bell rings if correct

The specific experiment:

- 15 subjects who considered themselves clairvoyant
- Each made 500 guesses, total of 7500
- Of these, 2006 were correct
- Compare to $7500/4 = 1875$

How significant is this?

ESP: analysis

Total of 7500 trials.

- Each time: one of four outcomes
- Total number of correct guesses: 2006

Null hypothesis: The data comes from a coin of bias 0.25.

Assume the null is true.

- The total number of successes in 7500 trials is approximately normal with what mean and standard deviation?

$$\begin{aligned} \text{Mean} = 7500 \times 0.25 &= 1875 \\ \text{Stddev} = \sqrt{7500 \times 0.25 \times 0.75} &\approx 37 \end{aligned}$$

- The z statistic:

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} \approx \frac{2006 - 1875}{37} \approx 3.5$$

This is strong evidence against the null.

Example: improving math scores?

National Assessment of Educational Progress data on 17-year olds:

- Average math score in 1978 was 300.4, with standard deviation 30.1
- Average math score in 1992 was 306.7, with standard deviation 34.9
- Both based on random sample of 1000 students

How significant was the improvement?

Null hypothesis: The means of the two distributions (scores in 1978, scores in 1992) are the same.

Assume the null is true. Let μ be the common mean.

- The sample average in 1978, call it X_1, is roughly normal with mean μ and standard deviation $\sigma_1 = 30.1/\sqrt{1000} \approx 1.0$.
- The sample average in 1992, call it X_2, is roughly normal with mean μ and standard deviation $\sigma_2 = 34.9/\sqrt{1000} \approx 1.1$.
- The difference $X_2 - X_1$ is therefore normally distributed with mean zero and standard deviation $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} \approx 1.5$.

What is the z-statistic? And what can we conclude?

Math scores, cont'd

100 students chosen at random in 1978 and 1992. Math scores recorded.

$$\begin{aligned} X_1 &= \text{sample average score in 1978} \\ X_2 &= \text{sample average score in 1992} \end{aligned}$$

Null hypothesis: The means of the two distributions (scores in 1978, scores in 1992) are the same.

Under the null, $X_2 - X_1$ is normally distributed with mean zero and standard deviation $\sigma = 1.5$.

Observed scores: $X_1 = 300.4$ and $X_2 = 306.7$.
The z-statistic for $X_2 - X_1$ is

$$z = \frac{(\text{observed}) - (\text{expected})}{\text{standard deviation}} = \frac{306.7 - 300.4}{1.5} \approx 2.1.$$

The observed difference has probability about 2% under the null: strong evidence against the null.

xample: the influence of wording

Study by Amos Tversky. 167 doctors were given information about the effectiveness of *surgery* versus *radiation therapy* for lung cancer. The same information was presented two ways.

80 of the doctors got Form A:

> *Of 100 people having surgery, 10 will die during treatment, 32 will have died by one year, and 66 will have died by five years. Of 100 people having radiation therapy, none will die during treatment, 23 will die by one year, and 78 will die by five years.*

The other 87 doctors got Form B:

> *Of 100 people having surgery, 90 will survive the treatment, 68 will survive one year or longer, and 34 will survive five years or longer. Of 100 people having radiation therapy, all will survive the treatment, 77 will survive one year or longer, and 22 will survive five years or longer.*

At the end, each doctor was asked which therapy he or she would recommend for a lung cancer patient.

	Form A	Form B
Favored surgery	40	73
Favored radiation	40	14
Total	80	87
Fraction favoring surgery	0.50	0.84

Let p_A be the probability that a doctor reading form A favors surgery, and let p_B be the probability that a doctor reading form B favors surgery. **Null hypothesis:** $p_A = p_B$.

Let X_A, X_B be the observed fractions favoring surgery.

- X_A is (roughly) normally distributed, with mean p_A and standard deviation $\sigma_A = \sqrt{(0.5 \times 0.5)/80} \approx 0.056$.
- X_B is (roughly) normally distributed, with mean p_B and standard deviation $\sigma_B = \sqrt{(0.84 \times 0.16)/87} \approx 0.039$.
- Under the null, $X_A - X_B$ is normally distributed with mean zero and standard deviation $\sigma = \sqrt{\sigma_A^2 + \sigma_B^2} \approx 0.068$.

Then $z \approx 5.0$. Very unlikely under the null!

ack to the Salk vaccine

	Size	Number of cases
Treatment	200,000	57
Control	200,000	142
No consent	350,000	92

Null hypothesis: Both groups have the same chance of getting polio.

Let X_t be the number of observed cases in the treatment group and X_c the number of observed cases in the control group.

- X_t is (roughly) normally distributed, with standard deviation $\approx \sqrt{57}$
- X_c is (roughly) normally distributed, with standard deviation $\approx \sqrt{142}$
- Under the null, $X_c - X_t$ is normally distributed with mean zero and standard deviation $\sqrt{57 + 142} \approx 14$.

The z statistic for $X_c - X_t$ is then

$$z \approx \frac{142 - 57}{14} \approx 6.1.$$

The observed difference is extremely unlikely under the null.

Statistical hypothesis testing

1. The z statistic
   - Testing the mean of a distribution
   - Testing whether two distributions have the same mean
2. The χ^2 statistic
   - Testing whether a sequence of $\{1, 2, \ldots, k\}$ outcomes comes from a particular k-sided die
   - Testing the independence of two variables

Testing a k-sided die

We have used the z-statistic to:

- Test whether the mean of a distribution is a certain value.
- Test whether two distributions have the same mean.

Eg. Checking whether a coin is fair.

But what if we want to check whether a k-sided die is fair?

- Rather like checking k different means, one for each outcome.
- Or, more precisely, $k-1$ different means.
- Could run $k-1$ separate tests.

Instead: run a single combined test with the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{((\text{observed frequency of } i) - (\text{expected frequency of } i))^2}{(\text{expected frequency of } i)}$$

and compare it to χ^2 distribution with $k-1$ degrees of freedom.

Example: is a die fair?

A gambler is concerned that the casino's die is loaded. He observes the following frequencies in a sequence of 60 tosses:

Outcome	1	2	3	4	5	6
Observed	4	6	17	16	8	9
Expected	10	10	10	10	10	10

Null hypothesis: die is fair.

Compute the χ^2 statistic for this data:

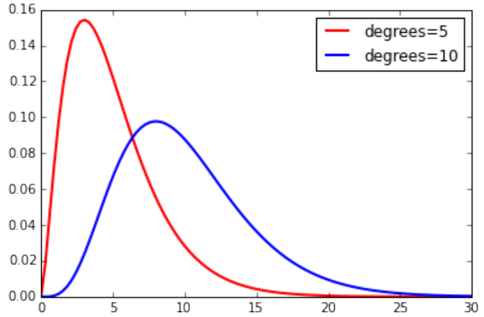
$$\chi^2 = \sum_{i=1}^{k} \frac{((\text{observed frequency of } i) - (\text{expected frequency of } i))^2}{(\text{expected frequency of } i)}$$

$$= \frac{6^2}{10} + \frac{4^2}{10} + \frac{7^2}{10} + \frac{6^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} \;=\; 14.2$$

Under the null, this value would be a random draw from a χ^2 distribution with 5 degrees of freedom.

Testing fairness of a die, cont'd

The χ^2 distribution:



The probability of getting a value as large as 14.2 (with 5 degrees of freedom) is 1.4%... strong evidence against the null.

Testing independence

Suppose there are k possible outcomes.

You have two sets of observations, $S_1, S_2 \subset \{1, 2, \ldots, k\}$.
Are they independent draws from the same distribution over $\{1, \ldots, k\}$?

- Null hypothesis: They are independent.
- Estimate the underlying distribution by combining the two samples. Call this P.
- Use the χ^2 statistic of how close S_1 and S_2 are to expected frequencies under P.

Example: left-handedness by sex

Data from a sample of 2,237 Americans of age 25-34:

	Men	Women
Right-handed	934 (87.5%)	1,070 (91.5%)
Left-handed	113 (10.6%)	92 (7.9%)
Ambidextrous	20 (1.9%)	8 (0.7%)

Is left-handedness really more common in men, or is this just a chance effect from sampling?

Null hypothesis: The two sets of numbers (for men and women) are independent draws from the same distribution.

Left-handedness, cont'd

Estimate the underlying distribution as well as expected frequencies for each of the two samples:

	Observed			Expected	
	Men	Women	Total	Men	Women
Right-handed	934	1,070	2,004 (89.6%)	956	1,048
Left-handed	113	92	205 (9.2%)	98	107
Ambidextrous	20	8	28 (1.2%)	13	15
Total	1,067	1,170	2,237	1,067	1,170

Compute the χ^2 statistic for this data:

$$\chi^2 = \sum_{\text{outcomes}} \frac{((\text{observed frequency}) - (\text{expected frequency}))^2}{(\text{expected frequency})}$$

$$= \frac{22^2}{956} + \frac{22^2}{1,048} + \frac{15^2}{98} + \frac{15^2}{107} + \frac{7^2}{13} + \frac{7^2}{15} \approx 12$$

Under the null, this would have a χ^2 distribution with 2 degrees of freedom. A value ≥ 12 has probability roughly 0.2%.