Sampling

DSE 210

Outline

1. Laws of large numbers
2. Basic sampling designs
3. Confidence intervals

Review: expected value

The expected value of a random variable X is

$$\mathbb{E}(X) = \sum_x x \Pr(X = x).$$

Example: A coin has heads probability p. Let X be 1 if heads, 0 if tails.

$$\mathbb{E}(X) \;=\; 1 \cdot p + 0 \cdot (1-p) \;=\; p.$$

Linearity properties:

- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ for any random variable X and any constants a, b.
- $\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_k)$ for any random variables $X_1, X_2, \ldots, X_k$.

Example: Toss n coins of bias p, and let X be the number of heads. What is $\mathbb{E}(X)$?

Let the individual coins be $X_1, \ldots, X_n$.

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np.$$

Review: variance

$\mathrm{var}(X) = \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2) - \mu^2$, where $\mu = \mathbb{E}(X)$.

Toss a coin of bias p. Let $X \in \{0, 1\}$ be the outcome.

$$\begin{aligned}
\mathbb{E}(X) &= p \\
\mathbb{E}(X^2) &= p \\
\mathbb{E}(X - \mu)^2 &= p^2 \cdot (1-p) + (1-p)^2 \cdot p = p(1-p) \\
\mathbb{E}(X^2) - \mu^2 &= p - p^2 = p(1-p)
\end{aligned}$$

This variance is highest when $p = 1/2$ (fair coin).

The standard deviation of X is $\sqrt{\mathrm{var}(X)}$.
It is the average amount by which X differs from its mean.

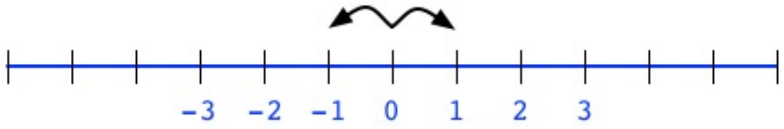
Useful variance rules:

- $\mathrm{var}(X_1 + \cdots + X_k) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_k)$ if X_i's independent.
- $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$.

Variance of a sum

$\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if the X_i are independent.

Symmetric random walk. A drunken man sets out from a bar. At each time step, he either moves one step to the right or one step to the left, with equal probabilities. Roughly where is he after n steps?



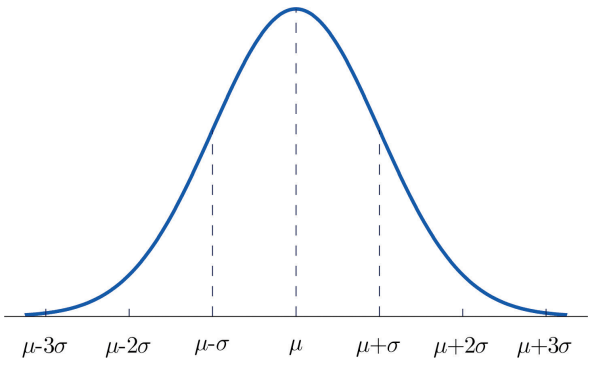
Let $X_i \in \{-1, 1\}$ be his ith step. Then $\mathbb{E}(X_i) = ?0$ and $\text{var}(X_i) = ?1$.

His position after n steps is $X = X_1 + \cdots + X_n$.

$$\mathbb{E}(X) = 0$$
$$\text{var}(X) = n$$
$$\text{stddev}(X) = \sqrt{n}$$

What is the distribution over his possible positions?
Approximately $N(0, n)$: Gaussian with mean 0 and std deviation $\sqrt{n}$.

The normal distribution



The normal (or *Gaussian*) $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- 68.3% of the distribution lies within one standard deviation of the mean, i.e. in the range $\mu \pm \sigma$
- 95.4% lies within $\mu \pm 2\sigma$
- 99.7% lies within $\mu \pm 3\sigma$

The central limit theorem

Suppose $X_1, \ldots, X_n$ are independent, and that they all come from the same distribution, with mean μ and variance σ^2.

Let $S_n = X_1 + \cdots + X_n$. Then S_n has mean and variance:

$$\mathbb{E}S_n = n\mu, \quad \text{var}(S_n) = n\sigma^2.$$

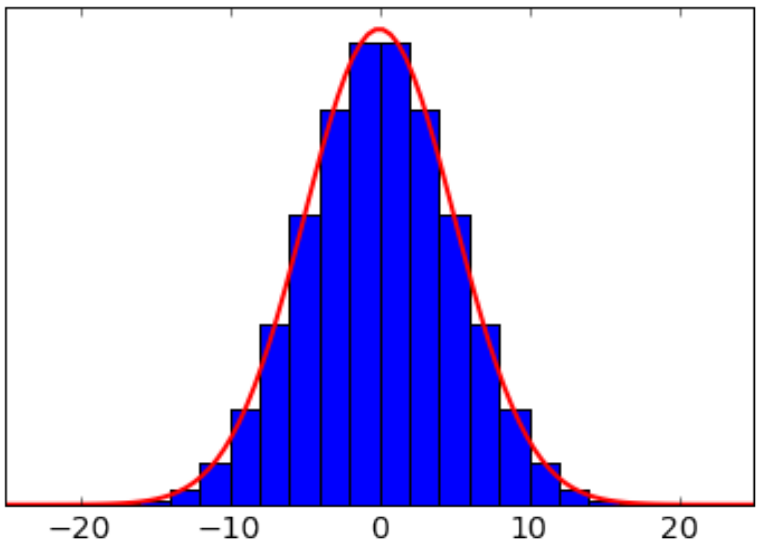
Central limit theorem, very roughly: For reasonably large n, the distribution of $S_n = X_1 + \cdots + X_n$ looks like $N(n\mu, n\sigma^2)$, the Gaussian with mean $n\mu$ and variance $n\sigma^2$.

Question: What does this imply about the average $(X_1 + \cdots + X_n)/n$? What does its distribution look like?

Answer: $N(\mu, \sigma^2/n)$.

Symmetric random walk, again

Each X_i is either 1 or -1, each with probability $1/2$.
Therefore, $X_1 + \cdots + X_n$ is distributed like $N(0, n)$.



25 steps

Tosses of a biased coin

A coin of bias (heads probability) p is tossed n times.

- What is the distribution of the observed **number** of heads, roughly?
  Answer: $N(np, np(1-p))$
  Mean np, standard deviation on the order of $\sqrt{n}$.
- What is the distribution of the observed **fraction** of heads, roughly?
  Answer: $N(p, p(1-p)/n)$.
  Mean p, standard deviation on the order of $1/\sqrt{n}$.

Example: A town has 30,000 registered voters, of whom 12,000 are Democrats. A random sample of 1,000 voters is chosen. How many of them would we expect to be Democrats, roughly?

Answer: The number of Democrats observed will roughly follow a $N(1000 \times 0.4, 1000 \times 0.4 \times 0.6) = N(400, 240)$ distribution.
This has mean 400 and standard deviation ≈ 15.5.

Outline

Sampling design

In the 1948 Presidential election, the polls all predicted Thomas Dewey as the winner, with at least a five-point margin. But the outcome was quite different.

Selection bias

The Republican bias in the Gallup Poll, 1936-1948.

Year	Gallup's prediction of Republican vote	Actual Republican vote
1936	44	38
1940	48	45
1944	48	46
1948	50	45

The safest way to sample is **at random**.

Multistage cluster sampling

Sometimes random sampling is inconvenient, and careful multistage procedures need to be used.

For instance,

1. Stage 1
   - Divide the US into four geographical regions: Northeast, South, Midwest, West.
   - Within each region, group together all population centers of similar sizes. E.g. All towns in the northeast with 50-250 thousand people.
   - Pick a random sample of these towns.
2. Stage 2
   - Divide each town into wards, and each ward into precincts.
   - Select some wards at random from the towns chosen earlier.
   - Select some precincts at random from among these wards.
   - Then select households at random from these precincts.
   - Then select members of the selected households at random, within the designated age ranges.

Sample size versus population size

A certain town in Illinois has the same balance of Democrats and Republicans as the nation at large. We want to determine these fractions using a random sample of 1000 people. Would it be better to choose the 1000 people from the town in Illinois, or from the entire country?

Let the unknown fraction be p. In both cases, the observed fraction will follow the $N(p, p(1-p)/1000)$ distribution.

What matters is the sample size, not the overall population size.

Outline

1. Laws of large numbers
2. Basic sampling designs
3. Confidence intervals

Example: estimating a fraction

A university has 25,000 registered students. In a survey, 400 students were chosen at random, and it turned out that 317 of them were living at home. Estimate the fraction of students living at home.

The observed fraction, out of $n = 400$ samples, is

$$\widehat{p} = \frac{317}{400} \approx 0.79.$$

Give error bars on this estimate.

Let p be the fraction of students living at home. Then:

$$\widehat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Therefore, $\widehat{p}$ has standard deviation $\sqrt{p(1-p)/n}$.

But we don't know p... so what error bar to use?

In a survey, $n = 400$ students were chosen at random, and it turned out that 317 of them were living at home.

The observed fraction living at home is $\widehat{p} = 0.79$. This value $\widehat{p}$ is normally distributed with mean p and standard deviation $\sqrt{p(1-p)/n}$.

Since we don't know the true standard deviation $\sqrt{p(1-p)}$ of each sample, use the observed standard deviation $\sqrt{\widehat{p}(1-\widehat{p})}$.

$$\text{stddev}(\widehat{p}) \approx \sqrt{\frac{0.79 \times 0.21}{400}} \approx 0.02.$$

Using normal approximation gives confidence intervals:

- 68.3% interval: 0.79 ± 0.02
- 95.5% interval: 0.79 ± 0.04
- 99.7% interval: 0.79 ± 0.06

What does a 95% confidence interval mean?
It means that if we were to do this over and over again, the interval would be correct (contain the true value) at least 95% of the time.

Estimating an average

In a certain town, a random sample is taken of 400 people age 25 and over. The average years of schooling of this sample is 11.6 years, with a standard deviation of 4.1. Find a 95% confidence interval for the average educational level of people 25 and over in this town.

What is the distribution of the observed average?

- Let the true mean educational level be μ, with stddev σ.
- We draw n samples from this distribution, and take the average $\widehat{\mu}$.
- This $\widehat{\mu}$ has distribution $N(\mu, \sigma^2/n)$.

Estimate the standard deviation of $\widehat{\mu}$.

- Its standard deviation is $\sigma/\sqrt{n}$.
- We don't know σ. Instead use the sample standard deviation, 4.1.
- Standard deviation of $\widehat{\mu}$ is roughly $4.1/\sqrt{400} \approx 0.2$.

Therefore, 95% confidence interval is 11.6 ± 0.4.
And recall: the chance is in the measuring procedure, not in the quantity being estimated.

DSE 210: Probability and statistics **Winter 2018**

Worksheet 11 — Sampling

1. A box contains 9 red marbles and 1 blue marbles. Nine hundred random draws are made from this box, with replacement. What is distribution of the number of red marbles seen, roughly?

2. Suppose that in the world at large, 1% of people are left-handed. A sample of 200 people is chosen at random. Give a 99% confidence interval for the number of them that are left-handed.

3. A dartboard is partitioned into 20 wedges of equal size, numbered 1 through 20. Half the wedges are painted red, and the other half are painted black. Suppose 100 darts are thrown at the board, and land at uniformly random locations on it.

   (a) Let X_i be the number of darts that fall in wedge i. What are $\mathbb{E}(X_i)$ and $\text{var}(X_i)$?

   (b) Using a normal approximation, give an upper bound on X_i that holds with 95% confidence.

   Let Z_r be the number of darts that fall on red wedges, let Z_b be the number of darts that fall on black wedges, and let $Z = |Z_r - Z_b|$ be the absolute value of their difference. We would like to get a 99% confidence interval for Z. To do this, define

   $$Y_i = \begin{cases} 1 & \text{if } i\text{th dart falls in red region} \\ -1 & \text{if } i\text{th dart falls in black region} \end{cases}$$

   and notice that $Z_r - Z_b$ can be written as $Y_1 + Y_2 + \cdots + Y_{100}$, the sum of independent random variables.

   (c) What are $\mathbb{E}(Y_i)$ and $\text{var}(Y_i)$?

   (d) Using the central limit theorem, we can assert that $Z_r - Z_b$ is approximately a normal distribution. What are the parameters of this distribution?

   (e) Give a 99% confidence interval for Z.

4. Suppose colorblindness appears in 1% of people. How large must a sample be in order for the probability of it containing at least one colorblind person to be at least 95%?

5. You have hired a polling agency to determine what fraction of San Diegans like sushi. Unknown to the agency, the actual fraction is exactly 0.5.

   The agency is going to poll a random subset of the population and return the observed fraction of sushi-lovers. How far off would you expect their estimate to be (i.e. what standard deviation) if:

   (a) they poll 100 people?

   (b) they poll 2500 people?

6. A sample is taken to find the fraction of females in a certain population. Find a sample size so that this fraction is estimated within 0.01 with confidence at least 99%.

7. A survey organization wants to take a simple random sample in order to estimate the percentage of people who have seen Downton Abbey. To keep the costs down, they want to take as small a sample as possible. But their client will only tolerate chance errors of 1% or so in the estimate. Should they use a sample of size 100, or 2500, or 10000? An auxiliary source of information suggests the population percentage will be in the range 20% to 40%.

8. In a certain city, there are 100,000 people age 18 to 24. A random sample of 500 of these people is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in the city who are enrolled in college. Give a 95.5% confidence interval for your estimate.

9. A survey research company uses random sampling to estimate the fraction of residents of Austin, Texas, who watch Spanish-language television. They are satisfied with the estimate they get using a sample size of 1,000 people.

   They then want to also estimate this fraction for Dallas, which has similar demographics to Austin, but twice the population. What sample size would be suitable for Dallas?

10. The National Assessment of Educational Progress tests nationwide samples of 17-year olds in school. In 1992, the students in a random sample of size 1000 averaged 307 on the math component of the test; the standard deviation of the scores was about 30.

    Estimate the nationwide average score on the math test. What is the standard deviation of this estimate?

11. A box contains many pieces of papers with numbers on them. 100 random draws are made from the box, with replacement, and the sum of the draws is 297.

    (a) Can you estimate the average of the numbers in the box?

    (b) Can you give a confidence interval for your estimate, based on the information so far?

12. A lake contains an unknown number of fish. 1000 of them are caught, marked with red spots, and then returned to the water. Later, a random subset of 100 fish are caught from the lake, and it is found that Z of them have red spots.

    (a) How would you estimate the number of fish in the lake, in terms of Z?

    (b) Let F be the true number of fish in the lake. In terms of F, give a normal approximation to the distribution of Z. with what parameters?

    (c) If you had to give a 95% confidence interval for the number of fish in the lake, what would it be?

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?

$$\text{Mean} = 7500 \times 0.25 \;=\; 1875$$
$$\text{Stddev} = \sqrt{7500 \times 0.25 \times 0.75} \;\approx\; 37$$

- 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.

$X_1 =$ sample average score in 1978

$X_2 =$ sample average score in 1992

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.

Example: 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!

Back 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.

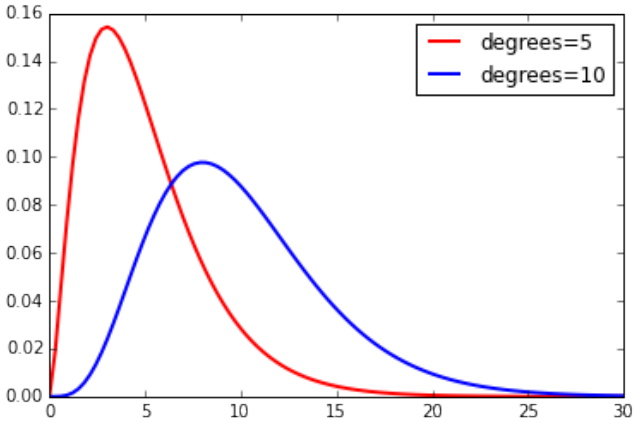
Compute the χ^2 statistic for this data:

$$\chi^2 = \sum_{i=1}^{k} \frac{((\text{observed frequency of } i) - (\text{expected frequency of } i))^2}{(\text{expected frequency of } i)}$$
$$= \frac{6^2}{10} + \frac{4^2}{10} + \frac{7^2}{10} + \frac{6^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} = 14.2$$

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%.

DSE 210: Probability and statistics **Winter 2018**

Worksheet 12 — Hypothesis testing

1. In the US in 1990, there were 2.1 million deaths from all causes, compared to 1.7 million in 1960: nearly a 25% increase. Does this data show that the public's health got worse over the period 1960–1990?

2. The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.

   (a) Was this a controlled experiment or an observational study?

   (b) Why did they study men and women and the different age groups separately?

   (c) The conclusion seems to be that you shouldn't start smoking, but once you've started, you shouldn't stop. Comment.

3. According to a study done by Kaiser Permanente in California, users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status.

   (a) Was this a controlled experiment or an observational study?

   (b) Why did the investigators adjust for age, education, and marital status?

   (c) The investigators concluded that the pill causes cervical cancer. Was this justified?

4. An experiment was carried out to determine the effect of providing free milk to school children in a certain district in Scotland. Some children in each school were chosen for the treatment group and got free milk; others were chosen as controls and got no milk. Assignment to the treatment or control groups was done at random, to make the two groups comparable in terms of family background and health.

   After randomization, teachers were allowed to use their judgement in switching children between treatment and control, to equalize the two groups. Was it wise to let teachers use their judgement in this way?

5. Studies of death certificates in the 1990s showed the average age of death was smaller for left-handed people than for right-handers.

   During the 20th century, there were big changes in child-rearing practices. In the early part of the century, parents insisted that their children be right-handed. By mid-century, parents were a lot more tolerant of left-handedness. Could this explain the observed discrepancy in average age at death of left- and right-handed people in the 1990s?

6. In 10,000 tossings, a coin came up heads 5,400 times. Should we conclude that the coin is biased?

   (a) Formulate the null hypothesis and alternative hypothesis.

   (b) Compute the z statistic and the p-value.

(c) What do you conclude?

7. A die is rolled 100 times. The total number of spots is 368 instead of the expected 350. Can this be explained as chance variation, or is the die loaded?

8. Other things being equal, which is better for the null hypothesis: a higher p-value or a lower p-value?

9. The National Household Survey on Drug Abuse was conducted in 1985 and 1992. In each year, a simple random sample of 700 people was used.

   (a) Among persons age 18 to 25, the percentage of marijuana users dropped from 21.9% to 11.0%. Is this difference real, or a chance variation?

   (b) Among persons age 18 to 25, the percentage of cigarette smokers dropped from 36.9% to 31.9%. Is this difference real, or a chance variation?

10. A random sample of 1000 freshmen at public universities were asked how many hours they worked each week (for pay). The average number of hours turned out to be 12.2, with a standard deviation of 10.5. A similar survey at private universities had an average of 9.2 hours, with a standard deviation of 9.9. Is the difference between these two averages due to chance?

11. John claims that he has extrasensory powers and can tell which of two symbols is on a card turned face down. To test his ability, he is asked to do a sequence of trials. The null hypothesis is that he is just guessing, so that the probability of being right on any trial is 1/2, whereas the alternative hypothesis is that he can name the symbol correctly more than half the time. Devise a test with the following properties:

    - If the null hypothesis is correct, then it is accepted at least 95% of the time.
    - If John can guess symbols correctly with probability $\geq 3/4$, then the alternative hypothesis is accepted at least 95% of the time.

12. A survey was conducted to determine the distribution of marital status by sex for persons age 25-29 in Wyoming. A random sample of 103 people was chosen, of whom 48 were men and 55 were women. The following results were obtained:

	Men	Women
Never married	43.8%	16.4%
Married	41.7%	70.9%
Widowed, divorced, separated	14.6%	12.7%

Are the distributions really different for men and women?