

Stat 88: Prob. & Mathematical Statistics in Data Science



<https://xkcd.com/892/>

Lecture 21 : 4/7/2022

Section 9.2, 9.3, 9.4

A/B testing & confidence intervals

Hypotheses tests: Review of steps

Say we want to see if a coin is fair, $p = 0.5$

For example: $H_0: p = \frac{1}{2}$
 $X = \# \text{ of H in } n \text{ tosses } X \sim \text{Bin}(n, p)$

1. State the **null hypothesis** (H_0) - that is, what is the assumption we are going to make. This will determine how we compute probabilities

2. State an **alternative hypothesis** (H_A) Note that this should not overlap with the null hypothesis, and it may or may not define probabilities (example: there was gender bias etc)

$$H_A \begin{cases} p \neq \frac{1}{2} \\ p < \frac{1}{2} \\ p > \frac{1}{2} \end{cases}$$

3. Decide on a **test statistic** to use that will help you decide which of the two hypotheses is supported by the evidence (data). Usually there is a natural choice. Use the null hypothesis to specify probabilities for the test statistic.

4. Find the **observed value** of the test statistic, and see if it is **consistent** with the null hypothesis. That is, compute the chance that we would see such an observed value, or more extreme values of the statistic (**p-value**).

T : test statistic $P(T \geq \text{obs value} \mid H_0 \text{ is true}) = \text{p-value}$
 for a right-tailed test

5. State your conclusion: whether you reject the null hypothesis or not. This is based on your chosen cutoff ("level of the test"). Reject if the **p-value** is less than the cutoff.

Suppose $H_A: p \neq \frac{1}{2}$, $\text{p-value} = P(T \geq \text{obs value} \mid H_0 \text{ is true})$


$H_A: p > \frac{1}{2}$

Example: Woburn



↑ obs. value of T

In the early 1990s, a leukemia cluster was identified in the Massachusetts town of Woburn. Many more cases of leukemia, a malignant cancer that originates in a bone marrow cell, appeared in this small town than would be predicted. Was this evidence of a problem in the town or just chance?

At the time US population was 280 mill. & # of Leukemia cases was 30,800 (in the US) $\rightarrow p = \frac{30800}{280 \times 10^6} \approx 0.00011$

Pop. of Woburn $\approx 35,000$

of Leukemia cases that year = 8 ← observed value of X
Assumption

where $X = \#$ of Leukemia cases $X \sim \text{Bin}(35000, 0.00011)$
Generic H_0 $E(X) \approx 3.85$

H_0 : Observed diff (b/w what you expect & what you see) is due to chance

① H_0 : There is no problem, we see 8 cases by chance & $X \sim \text{Bin}(35000, 0.00011)$
 $p = 0.00011$ or $np \approx 3.85$

② H_A : There is something going on, too many cases for it to be just chance.

③ Test statistic: X , den of X according to H_0
 $\text{Bin}(35000, 0.00011)$

④ observed value of X (denoted by x) = 8.

Going to use $\text{poisson}(3.85)$ to compute probs.

$$P(X \geq 8) = 1 - P(X < 8) = 1 - \text{stats.poisson.cdf}(7, 3.85) \\ \approx 0.0427 \leftarrow \text{assuming } H_0 \text{ is true.}$$

$$P(X \geq 8 \mid H_0 \text{ is true}) = 4.27\% \\ (\text{p-value})$$

If no cutoff specified for the exam, use 5%
(significance level)

Compare to 5% \rightarrow Reject the null

" " 1% \rightarrow Fail to reject the null.

Observed significance levels (a.k.a p -values)

- The p -value decides if observed values are *consistent* with the null hypothesis. It is a *tail* probability (also called *observed significance level*), and is the chance, **assuming that the null hypothesis is true**, of getting a test statistic equal to the one that was observed or even more in the direction of the alternative.
- If this probability is too small, then something is wrong, perhaps with your assumption (null hypothesis). That is, the data are unlikely if the null is true and therefore, your data are **inconsistent** with the null hypothesis.
- p -value is **not** the chance of null being true. The null is either true or not.
- The p -value is a *conditional probability* since it is computed *assuming* that the null hypothesis is true.
- The smaller the p -value, the stronger the evidence **against** the null and **towards** the alternative (in the direction of the alternative).
- Traditionally, below 5% ("result is statistically significant") and 1% ("result is highly significant") are what have been used. Significant means the p -value is small, not that the result is important.

Ex. 9.5.1

- All the patients at a doctor's office come in annually for a check-up when they are not ill. The temperatures of the patients at these check-ups are independent and identically distributed with unknown mean μ .
- The temperatures recorded in 100 check-ups have an average of 98.2 degrees and an SD of 1.5 degrees. Do these data support the hypothesis that the unknown mean μ is 98.6 degrees, commonly known as "normal" body temperature? Or do they indicate that μ is less than 98.6 degrees?

① $H_0: \mu = 98.6 \text{ deg}$

② $H_A: \mu < 98.6 \text{ deg.}$

③ Test statistic: \bar{X} (aka A_n)

④ Observed value = $\bar{x} = 98.2 \text{ deg}$

$$\frac{A_{100} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\frac{A_{100} - \mu}{\sigma/\sqrt{n}} = \frac{A_{100} - 98.6}{0.15}$$

X_1, X_2, \dots, X_{100} are the temperatures

Sample mean is your natural test statistic

Sample SD \approx pop. SD
(SD(X_k))

$$A_{100} = \bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \mu ??$$

$$\frac{\sigma}{\sqrt{n}} = \frac{1.5}{10} = 0.15$$

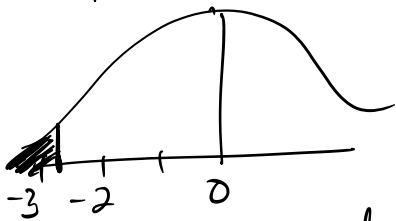
obs. value = 98.2

for normal

$$\begin{aligned} p\text{-value} &= P(A_{100} < 98.2 \mid H_0 \text{ is true}) \\ &= P\left(\underbrace{\frac{A_{100} - 98.6}{0.15}}_Z < \frac{98.2 - 98.6}{0.15} \mid \begin{array}{l} H_0 \text{ is true} \\ \text{so } A_{100} \sim N(98.6, 0.15) \end{array}\right) \end{aligned}$$

(Z is the usual notation for a std normal r.v.)

$$\begin{aligned} p\text{-value} &= P\left(Z < -\frac{0.4}{0.15}\right) = P(Z < -2.6667) \\ &\approx 0.0038 \\ &= 0.38\% \end{aligned}$$



Reject the null, it
doesn't seem likely that $\mu = 98.6$

A/B testing: comparing two distributions

- Data 8, section 12.3, randomized controlled trial to see if botulinum toxin could help manage chronic pain.
- 31 patients → 15 in treatment group, 16 in control group. 2 patients in the control group reported pain relief and 9 in the treatment group.
- A/B testing is a term used to describe hypothesis tests which involve comparing the distributions of two random samples. (Earlier we had one sample and made a hypothesis about its distribution.)
- In particular, we can conduct an A/B test for hypothesis tests involving results of randomized controlled trials, A is the control group and B the treatment group.

Fisher's exact test

\boxed{A}
 $n_A = 16$
 2

\boxed{B}
 $n_B = 15$
 9

- Control group: 16 patients, 2 reported relief
- Treatment group: 15 patients, 9 reported relief
- H_0 : The treatment has no effect (there would have been 11 patients reporting pain relief no matter what, and it just so happens that 9 of them were in the treatment group)
- H_A : The treatment has an effect

Test statistic: # of patients who had pain relief in sample

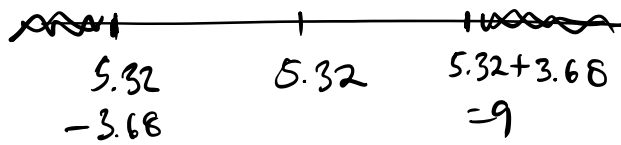
If H_0 is true, then the pop. of 31 has 20 Failures, 11 successes

$\begin{array}{|c|c|} \hline \text{Pain relief} & \text{no relief} \\ \hline \boxed{11} & \boxed{20} \\ \hline \end{array}$
 $N = 31$

$$H_0 \rightarrow X \sim \text{HG}(N=31, G=11, n=15) \quad E(X|H_0) = n \cdot \frac{G}{N} \approx 5.32$$

$$T = X - 5.32 = 9 - 5.32 = \underline{\underline{3.68}}$$

4/6/22 $p\text{-value} = P(X \leq 5.32 - 3.68) + P(X \geq 5.32 + 3.68)$



$$= P(X \leq 1.64) + P(X \geq 9)$$

$$= P(X \leq 1) + P(X \geq 9)$$

(b/c X is a whole #)

$$\approx \underline{0.00915}$$

using calc/computer

Example: The Lady Tasting Tea

- The first person to describe this sort of hypothesis test was the famous British statistician **Ronald Fisher**. In his book *The Design of Experiments*, he describes a tea party in which a lady of his acquaintance claimed that she could tell from tasting a cup of tea if the milk had been poured first or the tea.
- Fisher immediately set up an experiment in which she was given multiple cups of tea and asked to identify which of them had had the tea poured first. She tasted 8 cups of tea, of which 4 had the tea poured first, and identified 3 of them correctly. Does this data support her claim?

H_0 : The lady is guessing @ random

H_a : She can tell when tea is poured first

Actually happened.

Lady says	Actually happened.	
	Tea first	Milk first
Tea first	3	1
Milk first	1	3
		8

$X = \# \text{ of correct guesses}$
 $\# \text{ w/ tea poured first}$
 $\underline{N=8}, \underline{n=4}, G=4$

$X \sim HG(\overset{\text{tea poured first}}{\underline{N=8}}, \underline{G=4}, \underline{n=4})$

Observed value of $X = 3$ (in sample)

$$P(X \geq 3) \approx 0.243$$

Fail to reject the null. ~~Don't~~ conclude
 she is not guessing at random.

Example: Gender bias ?

- Rosen and Jerdee conducted several experiments using male bank supervisors (this was in 1974) who were given a personnel file and asked to decide whether to promote or hold the file. 24 were randomly assigned to a file labeled as that of a male employee and 24 to a female.
- 21 of the 24 males were promoted, and 14 of the females. Is there evidence of gender bias?

H_0 : No gender bias

H_A : There is a gender bias

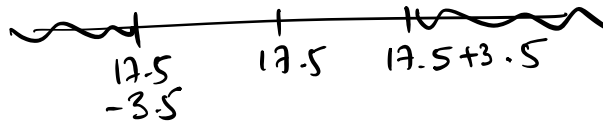
	M	F	
promoted	21	14	35
not promoted	3	10	13
	24	24	48

Let X = # of successes among male personnel files

$N = 48$, $G = 35$, $n = 24$ (sample of males)

$$X \sim \text{HG}(48, 35, 24), \quad E(X|H_0) = 24 \cdot \frac{35}{48} = 17.5$$

$$T = |X - 17.5| \quad \text{obs value} = 21 - 17.5 = 3.5$$


$$\begin{aligned} P(|X - 17.5| \geq 3.5) \\ &= P(X \leq 14) + P(X \geq 21) \\ &\approx 0.0489 \end{aligned}$$