Introduction to Hypothesis Testing

1. Is the die fair? What if the die is rolled 5 times and *three* comes up 4 times?

If the die is fair, what is the chance or probability of getting 4 "threes" in 5 rolls?

$n=5$, $p=\frac{1}{6}$, $k=4$ Repeated trials

$\binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right) = .0032$

4 or more "threes" $.0032 + .0001 \approx .0033$

2. Is the coin fair? How can you find out?

a) Suppose you toss it 10 times and observe 8 *heads*?

$n=10$, $p=\frac{1}{2}$ (assuming it is fair), $k=8$

$\binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 = \frac{45}{2^{10}} = .044$

Null hypothesis

Assume fair

b) Suppose you toss it 100 times and observe 60 *heads*?

Box Model → [0, 1] → Draw 100 & consider sum

Box AV $=\frac{1}{2}$, Box SD $=\frac{1}{2}$

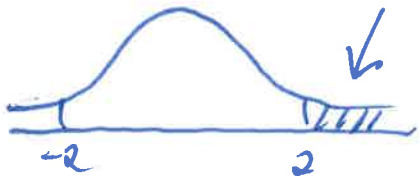
The sum of draws follows normal curve.

test statistic

The probability of 60 or more draws is found as follows:

EV for sum $= 50$, SE for sum $= \pm\sqrt{100}\cdot\frac{1}{2} = 5$

$\frac{60-50}{5} = 2$



$2\pm$%

p-value

3. One kind of plant has only blue flowers and white flowers. According to a genetic model, the offspring of a certain cross have a 75% chance to be blue-flowering, and a 25% chance to be white-flowering, independently of one another. Two hundred seeds of such a cross are raised, and 162 turn out to be blue-flowering. Is this consistent with the model? Answer yes or no, and explain briefly.

Null hypothesis

test statistic

Assume it is consistent.

Box Model: [1, 1, 1, 0] → Draw 200 & consider % is drawn.

The % is drawn follows normal curve

$$\text{Box SD} = \sqrt{\tfrac{3}{4}\cdot\tfrac{1}{4}} = \frac{\sqrt{3}}{4}$$

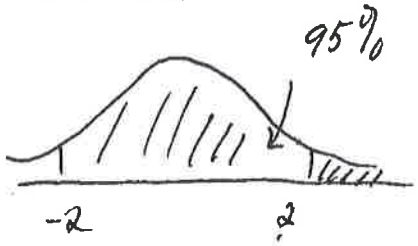
$$\text{EV for \%} = \frac{\text{Box SD}\cdot\sqrt{200}}{200}\times 100\% = \frac{\frac{\sqrt{3}}{4}\cdot\sqrt{200}}{200}\times 100\%$$

$$= 3.06\% \approx 3\%$$

$$\frac{162}{200}\times 100\% = 81\%$$

Z-test

$$\frac{81\% - 75\%}{3.06\%} \approx 2.$$



p-value is $2\frac{1}{2}\%$ or .025

There is strong statistical evidence against the null hypothesis. We reject; the evidence is not consistent with the model.

STATISTICAL TESTING

1. **Box Model:** Every legitimate test of significance involves a box model. The test gets at the question of whether an observed difference is real, or just a chance variation. A real difference is one that says something about the box, and doesn't just reflect a fluke of sampling.

2. **Null Hypothesis:** The null hypothesis says that an observed difference just reflects chance variation. The alternate hypothesis says that the observed difference is real.

3. **Test Statistic:** A test statistic measures the difference between the data and what is expected if the null hypothesis is true. You must be able to compute the probability values for the test statistic; that is, the test statistics must follow the normal curve (z-test), or a t-curve (t-test), or a chi-squared curve (χ^2 - test), or some other known probability variable.

4. **P-Value:** The probability value or the observed significance level of a test is the chance or probability of getting the test statistic as extreme as or more extreme than the observed one. The probability is computed on the basis that the null hypothesis is correct. The smaller this probability is, the stronger the evidence against the null hypothesis.

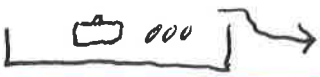
Federal Grand Jury

Probable Cause

Z-test

4. An airline company wishes to determine whether the average weight of suitcases carried by passengers between New York and London is more than 30 pounds. A sample of 64 is randomly selected and their suitcases weighed. The average of the sample was 32 pounds with an SD of 5 pounds. How significant is this difference?

Box Model



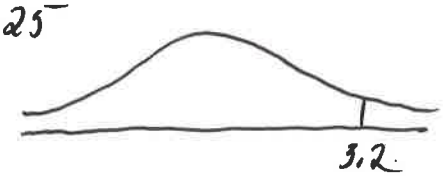
→ Draw 64 & consider AV of draws.

Null: Box AV = 30. AV of draws follows the normal curve

test statistic

EV = 30 SE for AV = $\frac{\text{Box SD} \cdot \sqrt{64}}{64} \approx \frac{5 \cdot \sqrt{64}}{64}$

$\frac{32-30}{.625} = 3.2$ $= .625$



p-value $< \frac{1}{2}\%$

Reject

5. Bottles of orange juice are supposed to have 16 fluid ounces. A random sample of 100 bottles from a large batch contains an average of 15.7 ounces with an SD of 0.2 ounces. Test the hypothesis that the bottles are being filled correctly, against the alternative that they are not full enough.

Box Model

→ Draw 100 & consider AV of draws

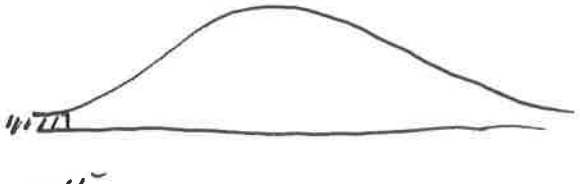
Null: Box AV = 16

Test statistic: AV of draws, follows normal curve

EV for AV = 16, SE for AV = Box SD $\times \sqrt{100} / 100$

$\approx \frac{(.2) \times 10}{100} = .02$

$\frac{15.7-16}{.02} = -15$



p-value ≈ 0

Reject!!

6. 400 people are each given a soda and a diet soda and asked to identify the diet soda. 283 correctly identify the diet soda. Is this evidence that people can tell the difference, or could they just be guessing?

1. Null hypothesis : No difference, cannot tell.
2. Box Model : [1, 0] → Draw 400 & consider the sum of draws.
3. Test statistic : The sum of the draws follows the normal curve. (Z-test)
4. Box AV = $\frac{1}{2}$, EV = 200
   Box SD = $\frac{1}{2}$ SE = $\frac{1}{2}\sqrt{400} = 10$

$$\frac{283 - 200}{10} = 8.3$$

p-value ≈ 0, Reject null.

7. **Notes on the P-value:**

We reject the null hypothesis if the **P-value** is small.

How small? Less than 5% is "**statistically significant**"
Less than 1% is "**highly statistically significant**"

The **P-value** is the chance of getting a ***sample value*** or ***test statistic*** at least as weird as the one we got, if the null hypothesis were true.

The **P-value** is called the "**observed significance level**".

The **P-value** is NOT the chance that the null hypothesis is true – it's the chance of us seeing DATA as far away as what we saw, if the null hypothesis *were* true. So if the **P-value** is small, we tend to believe the null is not true.

Two-tailed tests:

If you have a suspicion, *before you do the experiment* or take the sample, that the alternative hypothesis will be only in one direction, then do a 1-tailed test.

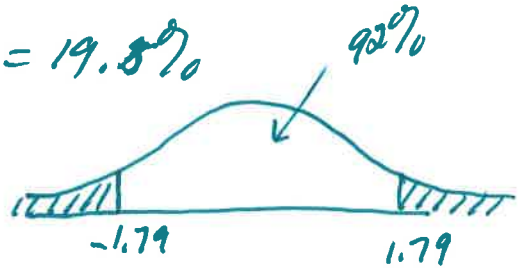
> Null: the average of the box is 25.
> Alt: the average is less than 25.

If you don't know in which direction it will go, then you should do a 2-tailed test.

> Null: the average of the box is 25.
> Alt: the average is NOT 25.

Example. National data suggest that 25% of Caucasians have a certain gene. To see whether Cache Valley people are similar to the nation with respect to this gene, a researcher takes a simple random sample of 200 people and finds that 39 of them have the gene. Is this evidence that Cache Valley people are different from the nation with respect to this gene?

1. Null : 25% of C.V. have the gene.
   Alternative : They don't, % ≠ 25%
2. [1, 0, 0, 0] → Draw 200 & consider % is drawn
   Box AV = $\frac{1}{4}$, Box SD = $\sqrt{\frac{1}{4}\cdot\frac{3}{4}} = \frac{\sqrt{3}}{4}$, EV for % = 25%
   SE for % = $\frac{\text{Box SD} \times \sqrt{200}}{200} \times 100\% = 3.06\%$
3. % of is drawn follows normal curve
4. p-value $\frac{39}{200} \times 100\% = 19.5\%$



$\frac{19.5 - 25}{3.06} = -1.79$

p-value = 8%.

The t-test:

The t-test is used when

- The number of draws is small (≤ 25).
- The numbers in the box follow the normal curve reasonable well.

It is similar to the z-test except

1. The box SD is approximated by the SD^+ of the sample.

   $$SD^+ = \sqrt{\frac{n}{n-1}}\ SD$$

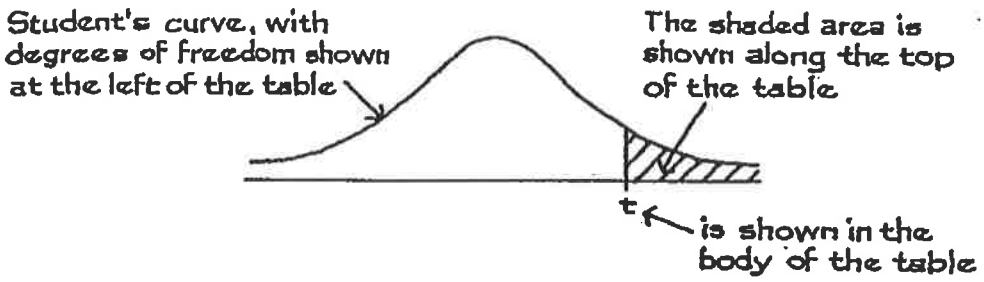
2. The test statistic is

   $$t = \frac{AV\ of\ draws\ -\ EV\ for\ AV\ of\ draws}{SE\ for\ AV\ of\ draws}$$

3. The degrees of freedom: $df = number\ of\ draws - 1$

4. Use the t-tables to get the p-value.

A *t*-TABLE



Degrees of freedom	*25%*	*10%*	*5%*	*2.5%*	*1%*	*0.5%*
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.81
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

Example: Toyota claims their new Prius hybrid car gets 51 miles per gallon on the highway. A random sample of 11 cars were tested and they averaged 48 mpg with and SD of 4.5 mpg. Is the difference significant?

1) Null hypothesis: Average mpg is 51.
Alternate: Average mpg < 51

2) [mpg's ...] → Draw 11 & consider the average of the draws.

3) Test statistic: $\frac{\text{AV of draws} - \text{EV for AV}}{\text{SE for AV}}$

follows t-curve, df = 10

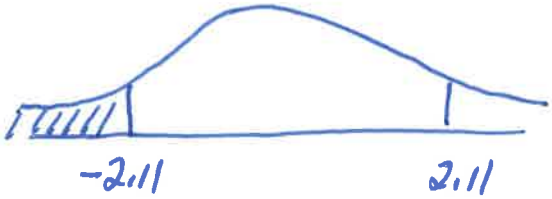
EV for AV = 51

SE for AV = $\frac{\text{Box SD} \times \sqrt{11}}{11}$

$\approx \frac{(\text{sample SD}^{+}) \times \sqrt{11}}{11} = \left[\sqrt{\frac{11}{10}}\,(4.5)\right]\frac{\sqrt{11}}{11}$

$= 1.42$

4) $\frac{48-51}{1.42} = -2.11$



p-value ≈ 2.5%

significant evidence against Toyota's claim.

Example: An English exam is taken by 2000 students. The exam scores are known to follow the normal curve. The teacher says that the population average of all 2000 test scores is 75, but one of the students thinks the population average is actually lower. She takes a simple random sample of 9 students and finds they got the following scores:

63, 53, 84, 82, 35, 50, 68, 73, 92.

AV = 66.6
SD = 17.2

Test to determine whether the population average really is 75, against the alternative that the student is correct. You should clearly state the null and the alternative hypothesis, find a test statistic and an approximate P-value, and state your conclusions in everyday language.

1) Null: AV = 75
Alternate: AV < 75

2) test scores → Draw 9 + consider AV

sample $SD^+ = \sqrt{\frac{9}{8}} \times 17.2 = 1.06 \times 17.2$
$= 18.24$

3) $\frac{\text{AV of draws} - \text{EV for AV}}{\text{SE for AV}}$ follows t-curve, df = 8

4) EV = 75, SE for AV = $\frac{\text{box SD} \times \sqrt{9}}{9}$

$\approx \frac{(18.24)\sqrt{9}}{9} = 6.08$

$\frac{66.6 - 75}{6.08} = -1.38$

t, df = 8

p-value ≈ 10%
fail to reject

ACCESS EXCELLENCE
@ the national health museum
ABOUT BIOTECH

Gregor Mendel (1823-1884)

Seung Yon Rhee

The theories of heredity attributed to Gregor Mendel, based on his work with pea plants, is well known to any student of biology. But his work was so brilliant and unprecedented at the time it appeared that it took thirty-four years for the rest of the scientific community to catch up to it. The short monograph, "Experiments with Plant Hybrids," in which Mendel described how traits were inherited, has become one of the most enduring and influential publications in the history of science.

Mendel, the first person to trace the characteristics of successive generations of a living thing, was not a world-renowned scientist of his day. Rather, he was an Augustinian monk who taught natural science to high school students. He was the second child of Anton and Rosine Mendel, farmers in Brunn, Moravia. Mendel's brilliant performance at school as a youngster encouraged his family to support his pursuit of a higher education, but their resources were limited, so Mendel entered an Augustinian monastery, continuing his education and starting his teaching career.

Mendel's attraction to research was based on his love of nature. He was not only interested in plants, but also in meteorology and theories of evolution. Mendel often wondered how plants obtained atypical characteristics. On one of his frequent walks around the monastery, he found an atypical variety of an ornamental plant. He took it and planted it next to the typical variety. He grew their progeny side by side to see if there would be any approximation of the traits passed on to the next generation. This experiment was "designed to support or to illustrate Lamarck's views concerning the influence of environment upon plants." He found that the plants' respective offspring retained the essential traits of the parents, and therefore were not influenced by the environment. This simple test gave birth to the idea of heredity.

Mendel's research reflected his personality. Once he crossed peas and mice of different varieties "for the fun of the thing," and the phenomena of dominance and segregation "forced themselves upon notice." He saw that the traits were inherited in certain numerical ratios. He then came up with the idea of dominance and segregation of genes and set out to test it in peas. It took seven years to cross and score the plants to the thousand to prove the laws of inheritance! From his studies, Mendel derived certain basic laws of heredity: hereditary factors do not combine, but are passed intact; each member of the parental generation transmits only half of its hereditary factors to each offspring (with certain factors "dominant" over others); and different offspring of the same parents receive different sets of hereditary factors. Mendel's work became the foundation for modern genetics.

The impact of genetic theory is no longer questioned in anyone's mind. Many diseases are known to be inherited, and pedigrees are typically traced to determine the probability of passing along an hereditary disease. Plants are now designed in laboratories to exhibit desired characteristics. The practical results of Mendel's research has not only changed the way we perceive the world, but also the way we live in it.

2. DID MENDEL'S FACTS FIT HIS MODEL?

Mendel's discovery ranks as one of the greatest in science. Today, his theory is amply proved and extremely powerful. But how good was his own experimental proof? Did Mendel's data prove his theory? Only too well, answered R. A. Fisher:

> ...the general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions. The data have evidently been sophisticated systematically, and after examining various possibilities, I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made.[7]

Leave the gardener aside for now. Fisher is saying that Mendel's data were fudged. The reason: Mendel's observed frequencies were uncomfortably close to his expected frequencies, much closer than ordinary chance variability would permit.

In one experiment, for instance, Mendel obtained 8,023 second-generation hybrid seeds. He expected $\frac{1}{4} \times 8{,}023 \approx 2{,}006$ of them to be green, and observed 2,001, for a discrepancy of 5. According to his own chance model, the data on seed color are like the results of drawing 8,023 times with replacement from the box

yellow	yellow	yellow	green

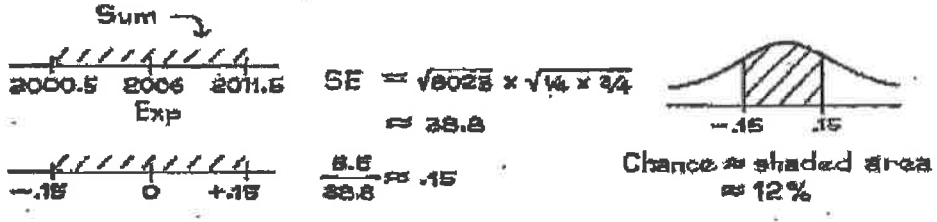
In this model, what is the chance of observing a discrepancy of 5 or less between the number of greens and the expected number? In other words, what is the probability that the number of greens will be

$$\text{between } \tfrac{1}{4} \times 8{,}023 - 5 \approx 2{,}001 \text{ and } \tfrac{1}{4} \times 8{,}023 + 5 \approx 2{,}011?$$

That is like drawing 8,023 times with replacement from the box

0	0	0	1

and asking for the chance that the sum will be between 2,001 and 2,011 inclusive. This chance can be estimated using the normal approximation, keeping track of the edges of the rectangles, as on p. 317.



About 88% of the time, chance variation would cause a discrepancy between Mendel's expectations and his observations greater than the one he reported.

By itself, this evidence is not very strong. The trouble is, every one of Mendel's experiments (with an exception to be discussed below) shows this kind of unusually close agreement between expectations and observations. Using the χ^2-test to pool the results (chapter 28), Fisher showed that the chance of agreement as close as that reported by Mendel is about four in a hundred thousand. To put this another way, suppose millions of scientists were busily repeating Mendel's experiments. For each scientist, imagine measuring the discrepancy between his observed frequencies and the expected frequencies by the χ^2-statistic. Then by the laws of chance, about 99,996 out of every 100,000 of these imaginary scientists would report a discrepancy between observations and expectations greater than the one reported by Mendel. That leaves two possibilities:

- either Mendel's data were massaged
- or he was pretty lucky.

The first is easier to believe.