• The following article appeared in the *Lafayette Journal & Courier* on November 30, 1974:

Praying to soybeans aids in higher yields

By George Cornell

With county officials measuring the results, experimenters on an Ohio farm say they found that portions of a field that had been the object of loving prayers yielded the biggest crop.

The case offered an unusual instance of recent stepped-up interest in psychic phenomena, viewed by many with keen skepticism.

"Somehow God's creative energy of growth can be channeled through us even to plants," says Gus Alexander of Wright State University, who holds a doctorate in communications research and who set up project.

It was carried out on a soybean field near Jamestown, Ohio, east of Dayton, with daily prayerful attention of a church group focused on six designated plots, but not on six adjoining control plots.

In checking the results of the experiment late in October, the Greene county agent's technical assistant, Donald H. Tate, was on hand to weigh the yields from the six experimental plots and six control plots.

According to the figures, five of the experimental strips had produced heavier yields than the adjacent control strips, while in the sixth case the control strip had a slightly greater yield.

In the experiment, a group of 10 people at Dayton's Church of the Golden Key, supplied with diagrams of the soybean plots, took on the task of "sending love" to the experimental areas each night at 11:30 p.m. for about 40 days.

SIGNIFICANCE TESTS

• *Significance tests* deal with the question of whether an observed result is "real," or just chance variation.

• Example *Praying to soybeans*.

• The experiment:

  • 12 plots

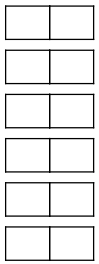
    • 6 treated with prayers

    • 6 controls

  • Assume the experiment was well-designed:

    • Random allocation to treatment and control.

    • "Blocking" — plots grouped into pairs, matched up on important variables, e.g., fertility, sunlight, drainage.



For each pair, a coin was tossed to decide which plot in the pair would receive prayers.

    • Blinding: Field workers kept ignorant of which plots were under treatment.

• Result: in 5 out of the 6 pairs of plots

T yield exceeded C yield.

- Possible explanations:
  - *Null hypothesis*:
    - Prayers have no effect on soybeans
    - Result here just reflects chance variation in the coin tosses.
    - Nothing interesting happened in this experiment.
  - *Alternative hypothesis*:
    - Soybeans really do respond to prayers.
  - No other explanations are reasonable.
    - The experiment was well-designed and conducted well.

  • If the researchers could convince us that the null hypothesis was untenable in view of the experimental result, we'd have to believe in the alternative hypothesis.
    - We want compelling evidence against the null hypothesis.

- Box model for the chance variation:

  • According to the null hypothesis, the number of pairs of plots where the treatment yield exceeds the control yield is like the sum of _____ draws made at random _____ replacement from the box

  _____ means T yield > C yield
  _____ means T yield < C yield

  • If prayers have no effect, whether or not T yield exceeds C yield gets determined by a coin flip, and has a 50–50 chance of going each way.

• Measuring the difference between the data and what's expected on the basis of the null hypothesis:

  • We *observed* 5 pairs of plots where T yield exceeded C yield.

  • Assuming the null hypothesis is true, we would have have *expected* T yield to exceed C yield in about

  expected value for the sum of the draws
  = (number of draws) × (fraction of 1's in box)
  = _____ × _____ = 3

  pairs, give or take

  SE for the sum of the draws
  $= \sqrt{\text{number of draws}} \times$ (SD of the box)
  $= \sqrt{6} \times$ _____ = 1.22

  pairs.

  • We can measure of the difference between the data and what's expected under the null hypothesis by

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{\text{observed difference}}{\text{it's likely size}}.$$

- The bigger this *z-statistic* is
  - the less consistent the experimental result is with the null, and the more consistent it is with the alternative.
  - the more the data deviate from what's expected under the null.
- For the experiment at hand, $z = (5 - 3)/1.22 = 1.63$.
  - How deviant is this result?

• Assessing the strength of the evidence against the null hypothesis — strategy:

  • Assuming the null hypothesis is true, work out the chance of getting a result as deviant, or more so, as the one actually observed.

  • The smaller is this *P-value*, the stronger is the evidence against the null hypothesis.

    • Suppose the P-value turns out to be 1/1000. That means that

    if the null hypothesis is true,

    only _____ in _____ similar experiments would give a result at least as extreme as the one in hand.

      • That's strong evidence against the null.

    • Suppose the P-value turns out to be 1/4. That means that

    if the null hypothesis is true,

    _____ out of _____ similar experiments would give a result at least as extreme as this one.

      • No reason to disbelieve the null.

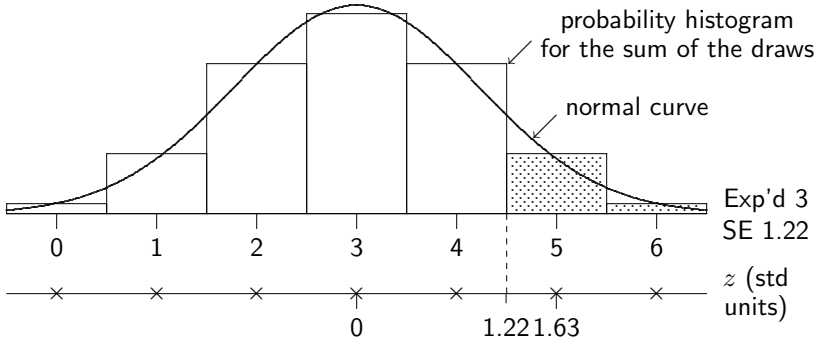
• The P-value is a measure of how surprising the observed effect is, if the null hypothesis is true. To put it another way, the P-value is a measure of how plausible the null hypothesis is, in the light of the observed effect. The _____ the P-value is, the harder it is to believe the null hypothesis is right.

• Assessing the strength of the evidence against the null hypothesis, continued:

  • Computation of the P-value.

    • We need the chance, computed under the null hypothesis, of getting a result at least as extreme as the one observed

    • That's the chance of getting a z at least as large as 1.63. This can be estimated using the normal approximation to the probability histogram for the sum of the draws:



      • The $\times$'s mark the possible values of z.

  • P-value = shaded area under the probability histogram
    $\approx$ area under the normal curve to the right of
    $= 1.22 \approx 11\%$.

    • Since the number of draws is small, it is important to keep track of the edges of the rectangles in the probability histogram when carrying out the normal approximation.

• Interesting, but not compelling, evidence against the null.

• Assessing the strength of the evidence against the null hypothesis, continued: If the experimenters had used 60 plots and found T yield exceeding C yield in 50 of them, the z-statistic would have been

$$\frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{50 - (60 \times 0.5)}{\sqrt{60}\sqrt{0.5 \times 0.5}} = \frac{20}{3.87} = 5.2,$$

corresponding to an extremely small P-value.

- This would have been very compelling evidence against the null.

• The experiment should have been run with more pairs of plots.

- Why didn't they subdivide the original 12 plots?

• Example. *Flex-time* is a time-scheduling technique used by some businesses.

- Employees are allowed to choose their working hours within broad limits set by management. This supposedly reduces absenteeism — days gone from work above and beyond vacations, etc.

• A company with 10,000 employees knows from past experience that over the last few years, absenteeism has averaged 6.3 days absent per year. To see if flex-time might change this, management puts all 10,000 employees on an experimental flex-time program for one year. Management chooses a simple random sample of 100 employees to follow in detail. At the end of the year, these employees

- average 5.4 days off from work, with an SD of 3.0 days.

• Possible explanations:

- Alternative hypothesis: flex-time made a real change in absenteeism.
- Null hypothesis: Average days off is still 6.3 days.
  - The lower sample average stems from chance variation in the sampling procedure. By the "luck of the draw," a few too many hard workers got into the sample.

• Box model: The sample is like _____ draws made at random _____ replacement from a box which has one ticket for each of the _____ employees, showing the _____ that employee was absent.

- Null hypothesis: the average of the box is $\mu_0 =$ ___.
- Alternative hypothesis: the average of the box is now some number $\mu \neq \mu_0$.

- Flex-time, continued.
- Test statistic:
  - If the null hypothesis is right, we'd expect days off in the sample to average to _____ days.
  - We observed a sample average of _____ days.
  - Observed − expected $= 5.4 - 6.3 = -0.9$ days.
  - As an estimate of the difference $\mu - \mu_0$ between μ and μ_0, -0.9 is likely to be off by about

    (estimated) SE of the average of the draws

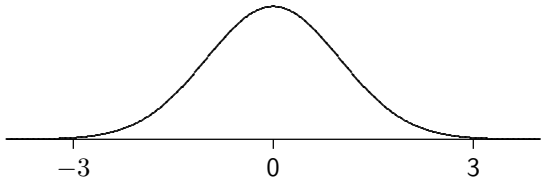
    $$= \frac{\text{estimated SD of box}}{\sqrt{\text{number of draws}}} = \frac{\quad}{\sqrt{100}} = 0.3 \text{ days}$$

- $z = \dfrac{\text{apparent difference}}{\text{likely contribution of chance}}$

  $$= \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{\text{signal}}{\text{noise}} = \frac{-0.9 \text{ days}}{0.3 \text{ days}} = -3$$

  - Think of the sample as "broadcasting" a message about the average of the box. The "signal" is that the average of the box is -0.9 days off from what's expected under the null. The signal is, however, distorted by "noise", chance variation in the sampling scheme. If the signal-to-noise ratio is large enough, then the message is "loud and clear" – the average of the box is not what the null hypothesis says it is.

- Flex-time, continued. $z = (\text{obs'd} - \text{exp'd})/\text{SE} = -3$.
- *P-value*
  - The alternative says that the average of the box is not 6.3 days. Deviations from 6.3, either up or down, favor the alternative over the null. In terms of z, both large positive and large negative values argue against the null.
  - If the null hypothesis is right, the probability histogram for the z-statistic is approximately _____ .
  - The chance of getting a value of the z-statistic at least as extreme — i.e., greater than 3, or less than -3 — as that actually observed is about _____%:



  - If the average days off per year hasn't changed from its past value of 6.3 days, only 1 in _____ studies similar to this one would have shown a greater apparent change.
- Conclusion:
  - Average days off per year really did change.
  - That's all the test of significance tells you. But management would want to know more — what?

- Flex-time, continued.
- How big was the change in average days off (for all 10,000 workers)?
  - Δ = (average for this year − average for past years) is estimated as -0.9 days, with a SE of 0.3 days.
  - 95%-confidence interval for Δ equals

$$\text{estimate} \pm 2 \text{ SE} = -0.9 \pm 2 \times 0.3 \text{ days}.$$

    - With 95% assurance, we can assert that absenteeism declined by somewhere between 0.3 days and 1.5 days.
- How important is the change in average days off?
  - That's question for management, not statistics.
- Did flex-time cause the change?
  - That's a question statistics can address.
  - This study was poorly designed, because absenteeism this year was compared to absenteeism in previous years. But this year may be different from the last (milder weather, more interesting work, etc.). So we can't tell if flex-time caused the change.
  - Can you suggest a better experimental design?

SIGNIFICANCE LEVELS

- The P-value is the chance, computing under the null hypothesis, of getting a value of the test statistic as extreme, or more so, than the observed one.
- True or false:
  - If the P-value is 43%, the null hypothesis looks plausible.
  - If the P-value is 0.43 of 1%, the null hypothesis looks plausible.
- A result whose P-value is less than α% is said to be *statistically significant at level* α%.
- A result which is statistically significant at the 5% level is called *statistically significant*.
  - In situations where the null hypothesis is right, chance variation produces a statistically significant result in about only 1 out of every 20 cases.
- A result which is statistically significant at the 1% level is called *highly statistically significant*.
  - In situations where the null hypothesis is right, chance variation produces a highly statistically significant result in about only 1 out of every 100 cases.
- A result which is (highly) statistically significant is one which is hard to attribute to chance variation, i.e., may be considered to be "real."

- True or false:
  - A "highly significant" result cannot possibly be due to chance.
    - _____. Even if the null hypothesis is true, 1% of the time the experiment will give a result which is "highly significant."
  - If a difference is "highly significant," there is less than a 1% chance for the null hypothesis to be right.
    - _____. A P-value does not give the chance of the null hypothesis being right. In fact, the P-value is computed on the basis of the null hypothesis.
- True or false:
  - If the P-value of a test is 2.9%, the result is statistically significant at the 3% level.
  - If the P-value of a test is 3.1%, the result is statistically significant at the 3% level.
  - The P-value of a test is the level at which the result is just statistically significant.
    - _____. For this reason, a P-value is often called the *observed significance level*.

ANALOGIES

- There are many analogies between hypothesis testing and criminal trials:

Jurisprudence	*Testing*
The defendant	The null hypothesis
The prosecuting attorney	The experimenter
The judge	The statistician
The jury	**S**cience
Guilty	Reject the null
Not guilty	Accept the null
Convicting the innocent	Rejecting the null when true. This is major egg in the face. You've proclaimed some result is real, but nobody can replicate your findings. A serious setback to your career.
Letting the guilty go free	Accepting the null when false. You failed to discover something that's really there. A disappointment to you, but not a setback to Science — since it's really there, somebody will find it.
Shadow of a doubt	Significance level
Beyond a shadow of a doubt	P-value $<$ significance level
The press	Journals

SUMMARY

• A *test of significance* gets at the question of whether an observed result is real (the _____ *hypothesis*) or just a chance variation (the _____ *hypothesis*).

• A legitimate test of significance requires a _____ _____ for the data. The null hypothesis has to be translated into a statement about the box model. Usually, the alternative does too.

• A _____ _____ is used to measure the difference between the data and what is expected under the _____ hypothesis. The *z-test* uses the statistic

$$z = \frac{\text{signal}}{\text{noise}} = \frac{\text{observed} - \text{expected}}{\text{SE}}.$$

The expected value and SE are computed on the basis of the _____ hypothesis.

• The _____ is the chance of getting a value for the test statistic as extreme as, or more so, than the observed one. The chance is computed on the basis that the _____ hypothesis is right. Therefore, the P-value does not give the chance of the null hypothesis being right.

• _____ P-values are evidence against the null hypothesis; they indicate that something besides _____ _____ was operating to produce the observed result.