Outline

Hypothesis testing

Philosophy: prove a claim by contradiction.

Analogy: "dependent love" story

Claim: You don’t love me.

Reasonning: If you loved me, you would take the trash out every week and put your socks away.

Data: Some weeks you don’t take the trash out or leave your socks where they fall.

Conclusion: You don’t love me.

Ingredients of a significance test

1. Null hypothesis H_0 and alternative hypothesis H_A.

   H_A = "you don't love me", H_0 = "you love me"

2. Statistic and null distribution

   summary data and what can be predicted under H_0

3. Measure evidence against H_0: p-value

   p-value measures the "compatibility" of data with H_0.

4. Make a decision and interpret in the context of the problem: accept H_A is small p-value, fail to reject H_0 if large p-value.

   adjust belief to H_A if data in contradiction with "you love me", or continue to believe H_0 if data consistent with "you love me"

The radiologists' missing sons

Male radiologists have long suspected that they tend to have fewer sons than daughters. Of 87 offspring of 'highly irradiated' radiologists, 30 were males. Assume this was a random sample, and proportion of male offspring is 0.51 in the human population.

We will try to prove the claim that $p \neq 0.51$ (H_A).

Proof by contradiction: determine what is expected if $p = 0.51$ (H_0) and whether the data is compatible with it.

p-value: measure of compatibility

The p-value is the probability, under H_0, of observing a result as extreme as or more extreme than that observed in the experiment.

The binomial test

1. $H_0 : p = 0.51$ versus $H_A : p \neq 0.51$
2. Summary statistic: $Y =$ # of sons among 87 offsprings.
   Null distribution: if H_0 is true, $Y \sim$
3. Under H_0 we expect to observe about 44.4 sons.
   Value actually observed: $y = 30$ sons.
   As extreme as 30 is: 30 or 59 (differ by 14.4 from 44.4)
   More extreme: $Y < 30$ or $Y > 59$.
   p-value $= 2 * \mathbb{P}\{Y \leq 30\} = 0.00277$
   Very tedious by hand but easy with R:

```
> sum(dbinom(0:30,size=87,p=.51))
[1] 0.001386064
> 2 * sum(dbinom(0:30,size=87,p=.51))
[1] 0.002772128
```

4. Data incompatible with H_0: reject H_0 and accept H_A.
   Strong evidence against H_0.

Significance tests: yes/no decision

Often the p-value is compared against $\alpha = 0.05$.
If p-value < 0.05, then we say **"Reject H_0 at the 5% level"** or **"The results are significant at the 5% level"**

The α-level needs to be set before seeing the data.
Common α-levels: 1%, 5%, and 10%.

The p-value measures incompatibility of data with H_0: interpreted as evidence against H_0. The smaller the p-value, the greater the evidence. Roughly:

$0.10 \leq p$	no evidence against H_0
$0.05 < p < 0.10$	weak evidence against H_0
$0.01 < p < 0.05$	moderate evidence against H_0
$0.001 < p < 0.01$	strong evidence against H_0
$p < 0.001$	very strong evidence against H_0

H_0 is either true or not true. Thus p-value $\neq \mathbb{P}\{H_0 \text{ is true}\}$.

Failing to reject H_0 is not proving H_0

We could repeat the test to

try to prove the claim H_A: $p \neq 0.35$.
We would fail to reject H_0: $p = 0.35$.
(recall data: $30/87 = 0.345$ sons)

try to prove the claim H_A: $p \neq 0.42$.
We would fail to reject H_0: $p = 0.42$.

Clearly we cannot accept both "$p = 0.35$" and "$p = 0.42$"

Interpretation: the **data is compatible** with "$p = 0.35$" as well as with "$p = 0.42$". Both are plausible values for p.
But very strong evidence that 0.51 is *not* the true value.

The binomial test: two-sided test with calculations

New example: 9 radiologists in experiment, 3 sons observed.

1. H_0: $p = 0.51$. Two-sided test: $H_A =$ "$p \neq 0.51$".
2. Statistic: # sons Y. Null distribution: $Y \sim \mathcal{B}(9, 0.51)$. Expectation: about 4.5 sons.
3. As extreme is 3 or 6 sons. More extreme is "$Y < 3$ or $Y > 6$ sons"

$$\begin{aligned} \mathbb{P}\{Y \leq 3 | p = 0.51\} &= \\ &= .001 + .01 + .06 + .15 = .23 \end{aligned}$$

so p-value $= 2\ \mathbb{P}\{Y \leq 3 | p = 0.51\} = 2 \times .23 = .47$

4. Conclusion: we fail to reject H_0, no evidence for H_A.

The binomial test: one-sided test, if prior evidence

1. H_0: $p = 0.51$. One-sided test: $H_A =$ "$p < 0.51$".
2. Statistic: # sons Y. Null distribution: $Y \sim \mathcal{B}(9, 0.51)$. Expectation: about 4.5 sons.
3. For this one-sided test: as or **more extreme in the direction of** H_A is "$Y \leq 3$ sons". So now

$$\text{p-value} = \mathbb{P}\{Y \leq 3 | p = 0.51\} = .23$$

4. Conclusion: we fail to reject H_0, no evidence for H_A.

The binomial test: one-sided test, if prior evidence

1. H_0: $p = 0.51$. One-sided test: $H_A =$ "$p > 0.51$".
2. Statistic: # sons Y. Null distribution: $Y \sim \mathcal{B}(9, 0.51)$. Expectation: about 4.5 sons.
3. For this one-sided test: as or **more extreme in the direction of** H_A is "$Y \geq 3$ sons". So now

$$\begin{aligned} \text{p-value} &= \mathbb{P}\{Y \geq 3 | p = 0.51\} = 1 - \mathbb{P}\{Y \leq 2\} \\ &= 1 - (\mathbb{P}\{Y = 0\} + \mathbb{P}\{Y = 1\} + \mathbb{P}\{Y = 2\}) \\ &= 1 - (.001 + 0.015 + 0.063) = .92 \end{aligned}$$

4. Conclusion: we fail to reject H_0, no evidence for H_A at all.

One-sided tests

We only perform one kind of test for a given experiment:
either one-sided: H_A: $p \neq p_0$
or two-sided: H_A: $p > p_0$, or H_A: $p < p_0$.

Most typically: two-sided test.

If a one-sided hypothesis H_A is used, it needs to be formulated **prior to seeing** the data and based on **prior evidence**.

The p-value is smaller with a one-sided test, unless the data goes in the opposite direction of a one-sided test: evidence for H_A is stronger with a one-sided test. Warning! resist the temptation!

Binomial test with R: binom.test()

```
> binom.test(30,87, p=.51)

        Exact binomial test

data:  30 and 87
number of successes = 30, number of trials = 87, p-value = 0.002488
alternative hypothesis: true probability of success is not equal to
                                                                0.51
95 percent confidence interval:
 0.2461396 0.4544136
sample estimates:
probability of success
             0.3448276

> binom.test(c(30,57), p=.51)

        Exact binomial test

data:  c(30, 57)
number of successes = 30, number of trials = 87, p-value = 0.002488
...
```

binom.test() in R - one-sided tests

```
> binom.test(30,87, p=0.51, alternative="less")

        Exact binomial test
data:  30 and 87
number of successes = 30, number of trials = 87, p-value = 0.001386
alternative hypothesis: true probability of success is less than 0.51
95 percent confidence interval:
 0.0000000 0.4374992
sample estimates:
probability of success
             0.3448276

> binom.test(30,87, p=0.51, alternative="greater")

        Exact binomial test
data:  30 and 87
number of successes = 30, number of trials = 87, p-value = 0.9993
alternative hypothesis: true probability of success is greater than 0.
95 percent confidence interval:
 0.2603165 1.0000000
sample estimates:
probability of success
             0.3448276
```

Outline

Testing a proportion with a z-test

Normal approximation to the binomial

If $X \sim \mathcal{B}(n, p)$ and if n is large enough so that both

$$np \geq 5 \quad \text{and} \quad n(1-p) \geq 5$$

(rule of thumb), then X and the sample proportion $\hat{p} = X/n$ are both approximately normally distributed:

$$X \sim \mathcal{N}(np, \sqrt{np(1-p)})$$
$$\hat{p} \sim \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}}) \qquad \text{approximately}$$

We can use this to avoid the heavy Binomial calculations and replace with easier Normal calculations:
The null distribution of Y is binomial, but sometimes can be approximated by a normal dist.

Testing a proportion with a z-test

Back to original data: 87 radiologists, 30 sons.

1. $H_0 : p = 0.51$ versus $H_A : p \neq 0.51$
2. **if H_0 is really true**, # of sons $Y \sim \mathcal{B}(87, 0.51)$, which $\approx$ some normal distribution, because $np = 87 * 0.51 = 44.37 > 5$ and $n(1-p) = 42.63 > 5$.

   $\mu_Y =$ $= 44.37$, $\sigma_Y =$ $= 4.66$.

   So $Y \sim \mathcal{N}(44.37, 4.66)$ approximately: null distribution.
3. 30 sons observed. As or more extreme: $Y \leq 30$ or $Y \geq 59$.

   p-value $= 2\,\mathbb{P}\{Y \leq 30 | p = 0.51\} \approx 2\,\mathbb{P}\left\{ \right\}$

   $=$ $= 0.002$.
4. Accept H_A at $\alpha = .005$ level. Strong evidence against H_0.

Testing a proportion with a z-test

Same test, but using the sample proportion: $\hat{p} = 30/87 = 0.345$

1. $H_0 : p = 0.51$ versus $H_A : p \neq 0.51$
2. **if H_0 is really true**, sample proportion $\hat{p}$ is approximately normally distributed because
   $np = 87 * 0.51 = 44.37 > 5$ and $n(1-p) = 42.63 > 5$.

   mean: $p = 0.51$, std. dev: $\sigma_{\hat{p}} =$ $= .0536$ so

   $$\hat{p} \sim \mathcal{N}(.51, .0536) \text{ approx.}$$

3. $\hat{p} = .345$ observed.

   p-value $= 2\,\mathbb{P}\{\hat{p} \leq .345 | p = 0.51\} \approx 2\,\mathbb{P}\Big\{$ $\Big\}$

   $=$ $= 0.002.$

4. Accept H_A at $\alpha = .005$ level. Strong evidence against H_0.

R commands for z-tests

Using the normal approximation for Y:

```
> 2*pnorm(30, mean=44.37, sd=4.663)
[1] 0.002058
```

Using the normal approximation for sample proportion $\hat{p}$:

```
> 2*pnorm(30/87, mean=.51, sd=.0536)
[1] 0.002059
```

Outline

Confidence intervals for proportions

What is the probability of getting the flu if one

has gotten the shot during the Fall,

and is in contact with the virus in the winter?

Experiment: Randomly sample $n = 37$ persons, get them the shot in the Fall. Expose them to the virus in December.
$Y = \#$ of persons in the experiment who get the disease (the shot didn't give them enough protection). We observe $y = 5$.

$p =$ true value in the population: proportion or probability.

$\hat{p} = Y/n$ observed value. Here $\hat{p} = 5/30 = 0.17$.

Goal: 95% confidence interval for p.

Confidence intervals for proportions

Recall distribution of $\hat{p}$:

Mean of $\hat{p}$: $\mu_{\hat{p}} = p$,

Std. dev., or standard error of $\hat{p}$: $\text{SE}_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$.

If large n (i.e $np \geq 5$ and $n(1-p) \geq 5$), then approximately normal distribution.

$\hat{p}$ lies in $p \pm 1.96\sqrt{\dfrac{p(1-p)}{n}}$ in about 95% of experiments, i.e.

p lies in $\hat{p} \pm 1.96\sqrt{\dfrac{p(1-p)}{n}}$ in about 95% of experiments.

Wald-type confidence intervals

First idea: plug-in $\hat{p}$ in place of p and use

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

as a 95% confidence interval.

Flu cases: $y = 5$ out of $n = 30$, so $\hat{p} = 5/30 = .17$ and $\sqrt{\frac{.17(1-.17)}{30}} = .07$. Wald-type 95% confidence interval:

$$.17 \pm 1.96 * .07 = .17 \pm .13 \quad \text{i.e. } (.033, .300)$$

BUT: this does not work very well. p lies in $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ in

84% of the experiments if $n = 10$ and $p = .3$,

65% of the experiments if $n = 10$ and $p = .1$,

We are over-estimating our confidence.

Agresti-Coull confidence intervals

Instead: We pretend we have 4 more observations (i.e. sample size is $n + 4$) and that out of those 4 extra observations, there are 2 successes and 2 failures (i.e. # successes is $Y + 2$).

$$\tilde{p} = \frac{y+2}{n+4} \qquad \text{and } \mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

A 95% confidence interval for p is

$$\tilde{p} \pm 1.96\ \mathrm{SE}_{\tilde{p}}$$

p lies in $\tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ in

- 95.2% of the experiments if $n = 10$ and $p = .3$,
- 93% of the experiments if $n = 10$ and $p = .1$,

We are no more over-estimating our confidence!

Agresti-Coull confidence intervals

Example: $n = 30$, $y = 5$ flu cases.

We get $\tilde{p} = (5 + 2)/(30 + 4) = .21$ and
$\text{SE}_{\tilde{p}} = \sqrt{.21 * .79/34} = .07$.

Our 95% confidence interval is $(0.070, .342)$.

How big should n be?

How many people should I sample so that my margin of error is at most 1% ?

margin or error $= 1.96 * \text{SE}$, so it means SE at most 0.5%, i.e $\text{SE}_{\tilde{p}} \leq 0.005$. But $\text{SE}_{\tilde{p}}$ is

$$\text{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} \leq \sqrt{\frac{1/4}{n+4}}$$

We then need (safe choice)

$$n = \frac{1}{4(\text{Desired SE})^2} - 4$$

Example: for SE at most 0.005, we need $n \geq 10,000 - 4$.
That's why polls are usually done on several thousands people.