Hypothesis Testing

Mark Lunt

Centre for Epidemiology Versus Arthritis
University of Manchester

01/11/2022

Introduction

- We saw last week that we can never know the population parameters without measuring the entire population.
- We can, however, make inferences about the population parameters from random samples.
- Last week, we saw how we can create a confidence interval, within which we are reasonably certain the population parameter lies.
- This week, we will see a different type of inference: is there evidence that the parameter **does not** take a particular value ?

Hypothesis Testing

- Form the Null Hypothesis
- Calculate probability of observing data if null hypothesis is true (p-value)
- Low p-value taken as evidence that null hypothesis is unlikely
- Originally, only intended as informal guide to strength of evidence against null hypothesis

Significance Testing

- Fisher's p-value was very informal way to assess evidence against null hypothesis
- Neyman and Pearson developed more formal approach: significance testing
- Based on decision making: rule for deciding whether or not to reject the null hypothesis
- Clinically, need to make decisions. Scientifically, may be more appropriate to retain uncertainty.
- Introduces concepts of power, significance

The Null Hypothesis

- Simplest acceptable model.
- If the null hypothesis is true, the world is uninteresting.
- Must be possible to express numerically (“test statistic”).
- Sampling distribution of test statistic must be known.

The Alternative Hypothesis

- "Null Hypothesis is untrue"
- Covers any other possibility.
- May be one-sided, if effect in opposite direction is as uninteresting as the null hypothesis

One and Two-sided tests

- Good example: χ^2 test.
  - χ^2 test measures difference between expected and observed frequencies
  - Only unusually large differences are evidence against null hypothesis.
- Bad example: clinical trial
  - A drug company may only be interested in how much better its drug is than the competition.
  - Easier to get a significant difference with a one-sided test.
  - The rest of the world is interested in differences in either direction, want to see a two-sided test.
- One-sided tests are rarely justified

Test Statistic

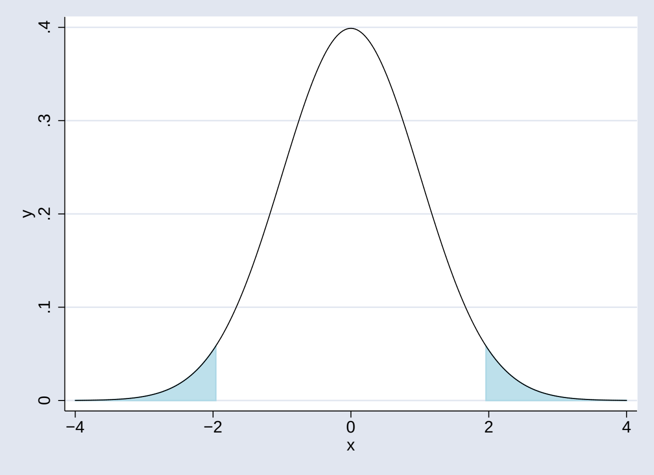
- Null hypothesis distribution must be known.
  - Expected value if null hypothesis is true.
  - Variation due to sampling error (standard error) if null hypothesis is true.
- From this distribution, probability of any given value can be calculated.
- Can be a mean, proportion, correlation coefficient, regression coefficient etc.

Normally Distributed Statistics

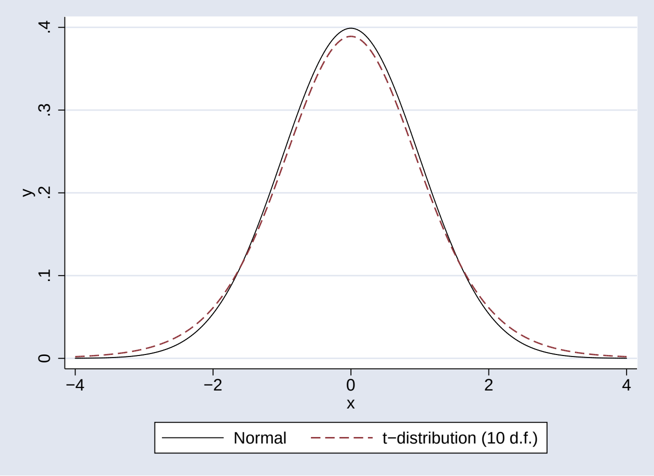
- Many test statistics can be considered normally distributed, if sample is large enough.
- If the test statistic T has mean μ and standard error σ, then $\frac{T-\mu}{\sigma}$ has a normal distribution with mean 0 and standard error 1.
- We do not know σ, we only have estimate s.
- If our sample is of size n, $\frac{T-\mu}{s}$ has a t-distribution with $n-1$ d.f.
- Hence the term "t-test".
- If $n \geq 100$, a normal distribution is indistinguishable from the t-distribution.
- Extreme values less unlikely with a t-distribution than a normal distribution.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Test statistic: Normal distribution



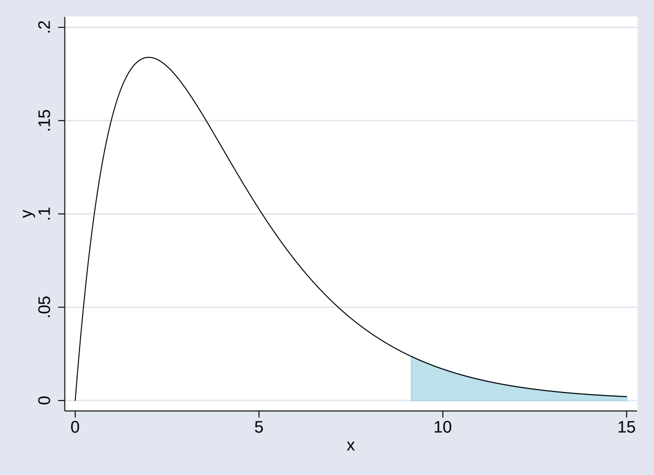
T-distribution and Normal Distribution



Non-Normally Distributed Statistics

- Statistics may follow a distribution other than the normal distribution.
  - χ^2
  - Mann-Whitney U
- Many will be normally distributed in large enough samples
- Tables can be used for small samples.
- Can be compared to quantiles of their own distribution

Test Statistic: χ^2_4



CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Example 1: Height and Gender

Null hypothesis On average, men and women are the same height

Alternative Hypothesis One gender tends to be taller than the other.

Test Statistic Difference in mean height between men and women.

One-Sided Hypotheses

- Men are taller than women
- Women are taller than men

Example 2: Drinks preferences

Null hypothesis Equal numbers of people prefer Coke and Pepsi

Alternative Hypothesis Most people prefer one drink to the other

Test Statistic Several possibilities:

- Difference in proportions preferring each drink
- Ratio of proportions preferring each drink

One-Sided Hypotheses

- More people prefer Coke
- More people prefer Pepsi

The *p*-value

- Probability of obtaining a value of the test statistic at least as extreme as that observed, *if the null hypothesis is true*.
- Small value $\Rightarrow$ data obtained was unlikely to have occurred under null hypothesis
- Data did occur, so null hypothesis is probably not true.
- Originally intended as informal way to measure strength of evidence against null hypothesis
- It it **not** the probability that the null hypothesis is true.

Interpreting the *p*-value

- $0 \leq p \leq 1$
- Large p (≥ 0.2, say) $\Rightarrow$ no evidence against null hypothesis
- $p \leq 0.05 \Rightarrow$ there is some evidence against null hypothesis
- Effect is “statistically significant at the 5% level”
- 0.05 is an arbitrary value: 0.045 is very little different from 0.055.
- Smaller $p \Rightarrow$ stronger evidence
- Large *p*-value not evidence that null hypothesis is true.

Factors Influencing *p*-value

- **Effect Size**: a big difference is easier to find than a small difference.
- **Sample Size**: The more subjects, the easier to find a difference
- Always report actual *p*-values, not $p < 0.05$ or $p > 0.05$
- *NS* is unforgivable
- “No significant difference” can mean “no difference in population” or “Sample size was too small to be certain”
- Statistically significant difference may not be clinically significant.

Interpreting a significance test

- Significance test asks “Is the null hypothesis true”
- Answers either “Probably not” or “No comment”
- Answers are interpreted as “No” or “Yes”
- Misinterpretation leads to incorrect conclusions

Meta-Analysis Example

- Ten studies
- 50 unexposed and 50 exposed in each
- Prevalence 10% in unexposed, 15% in exposed
- True RR = 1.5

Meta-Analysis Results

Study	RR	*p*-value
1	1.0	1.00
2	3.0	0.16
3	2.0	0.23
4	0.7	0.40
5	1.8	0.26
6	1.3	0.59
7	1.6	0.38
8	1.8	0.34
9	1.4	0.45
10	1.4	0.54

Meta-Analysis Results

Study	RR	*p*-value
1	1.0	1.00
2	3.0	0.16
3	2.0	0.23
4	0.7	0.40
5	1.8	0.26
6	1.3	0.59
7	1.6	0.38
8	1.8	0.34
9	1.4	0.45
10	1.4	0.54
Pooled Data	1.4	0.04

Inappropriate Hypothesis Testing

- Table 1 in a double-blind randomised controlled trial

- Identifying confounders in observational studies

Inappropriate Hypothesis Testing

- Table 1 in a double-blind randomised controlled trial
  - Null hypothesis: both arms are random samples from the same population
  - Null hypothesis is true by design
  - Small p-value: this is a rare event, but not really evidence of scientific malpractice
- Identifying confounders in observational studies

Inappropriate Hypothesis Testing

- Table 1 in a double-blind randomised controlled trial
  - Null hypothesis: both arms are random samples from the same population
  - Null hypothesis is true by design
  - Small *p*-value: this is a rare event, but not really evidence of scientific malpractice
- Identifying confounders in observational studies
  - Confounding effect depends on size of difference
  - *P*-value also depends on size of study
  - Can have same confounding and different $p-$values and vice versa

Getting it Wrong

- There are two ways to get it wrong:
  - The null hypothesis is true, we conclude that it isn't (Type I error).
  - The null hypothesis is not true, we conclude that it is (Type II error).

Type I Error (α)

- Null hypothesis is true, but there is evidence against it.
- 1 time in 20 that the null hypothesis is true, a statistically significant result at the 5% level will be obtained.
- The smaller the *p*-value, the less likely we are making a type I error.
- Testing several hypotheses at once increases the probability that at least one of them will be incorrectly found to be statistically significant.
- Several corrections are available for “Multiple Testing”, *Bonferroni’s* is the most commonly used, easiest and least accurate.
- Some debate about whether correction for multiple testing is necessary, but state how many tests were done.

Type II Error (β)

- Null hypothesis is not true, but no evidence against it in our sample.
- Depends on study size: small studies less likely to detect an effect than large ones
- Depends on effect size: large effects are easier to detect than small ones
- **Power** of a study = 1 - β = Probability of detecting a given effect, if it exists.

Testing $\bar{x}$

- Can compare $\bar{x}$ to a hypothetical value (e.g. 0).
- Sometimes called "One-sample t-test".
- Test statistic $T = \frac{\bar{x}-\mu}{S.E.(x)}$.
- Compare T to a t-distribution on $n-1$ d.f.

Testing $\bar{x}$: Example

- The following data are uterine weights (in mg) for a sample of 20 rats. Previous work suggests that the mean uterine weight for the stock from which the sample was drawn was 24mg. Does this sample confirm that suggestion ?
- 9, 14, 15, 15, 16, 18, 18, 19, 19, 20, 21, 22, 22, 24, 24, 26, 27, 29, 30, 32
- $\bar{x} = 21.0$
- S.D.$(x) = 5.912$

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Testing $\bar{x}$: Solution

$$
\begin{aligned}
S.E.(x) &= \frac{5.912}{\sqrt{20}} \\
&= 1.322
\end{aligned}
\qquad\qquad
\begin{aligned}
T &= \frac{\bar{x} - 24.0}{S.E.(x)} \\
&= \frac{21.0 - 24.0}{1.322} \\
&= -2.27
\end{aligned}
$$

- Comparing -2.27 to a t-distribution on 19 degrees of freedom gives a *p*-value of 0.035
- I.e if the stock had a mean uterine weight of 24mg, and we took repeated random samples, less than 4 times in 100 would a sample have such a low mean weight.

One-Sample t-test in Stata

```
. ttest x = 24

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      20          21    1.321881     5.91163    18.23327    23.76673
------------------------------------------------------------------------------
Degrees of freedom: 19

                              Ho: mean(x) = 24

     Ha: mean < 24               Ha: mean != 24                Ha: mean > 24
       t =  -2.2695                t =  -2.2695                t =  -2.2695
   P < t =   0.0175          P > |t| =   0.0351            P > t =   0.9825
```

The Unpaired (two-sample) T-Test

- For comparing two means
- If we are comparing x in a group of size n_x and y in a group of size n_y,
  - Null hypothesis is $\bar{x} = \bar{y}$
  - Alternative hypothesis is $\bar{x} \neq \bar{y}$
  - Test statistic
$$T = \frac{\bar{y} - \bar{x}}{\text{S.E. of } (\bar{y} - \bar{x})}$$
  - T is compared to a t distribution on $n_x + n_y - 2$ degrees of freedom
- You may need to test (`sdtest`) whether the standard deviation is the same in the two groups.
- If not, use the option `unequal`.

Two-Sample t-test in Stata

```
. ttest nurseht, by(sex)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  female |     227     159.774    .4247034    6.398803    158.9371    160.6109
    male |     175    172.9571    .5224808    6.911771    171.9259    173.9884
---------+--------------------------------------------------------------------
combined |     402    165.5129    .4642267    9.307717    164.6003    166.4256
---------+--------------------------------------------------------------------
    diff |           -13.18313    .6666327               -14.49368   -11.87259
------------------------------------------------------------------------------
Degrees of freedom: 400

                 Ho: mean(female) - mean(male) = diff = 0

     Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
       t = -19.7757               t = -19.7757               t = -19.7757
   P < t =    0.0000          P > |t| =    0.0000          P > t =    1.0000
```

Comparing Proportions

- We wish to compare $p_1 = \frac{a}{n_1}$, $p_2 = \frac{b}{n_2}$
- Null hypothesis: $\pi_1 = \pi_2 = \pi$
- Standard error of $p_1 - p_2 = \sqrt{\pi(1 - \pi)(\frac{1}{n_1} + \frac{1}{n_2})}$
- Estimate π by $p = \frac{a+b}{n_1+n_2}$
- $\frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$ can be compared to a standard normal distribution

Comparing Proportions in Stata

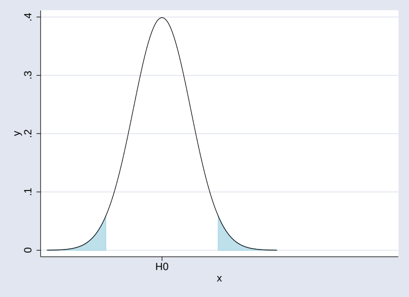
```
. cs back_p sex

                 | sex                    |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |       637         445  |       1082
        Noncases |      1694        1739  |       3433
-----------------+------------------------+-----------
           Total |      2331        2184  |       4515
                 |                        |
            Risk |  .2732733    .2037546  |   .2396456
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .0695187       |    .044767    .0942704
      Risk ratio |         1.341188       |   1.206183    1.491304
 Attr. frac. ex. |         .2543926       |   .1709386     .329446
 Attr. frac. pop |         .1497672       |
                 +-------------------------------------------------
                               chi2(1) =    29.91  Pr>chi2 = 0.0000
```

Sample Size

- Given:
  - Null hypothesis value
  - Alternative hypothesis value
  - Standard error
  - Significance level (generally 5%)
- Calculate:
  - Power to reject null hypothesis for given sample size
  - Sample size to give chosen power to reject null hypothesis

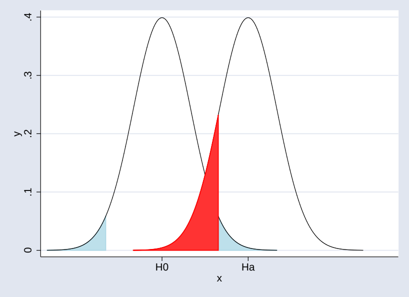
Power Calculations Illustrated: 1



Shaded area $=$ Sample value significantly different from H_0

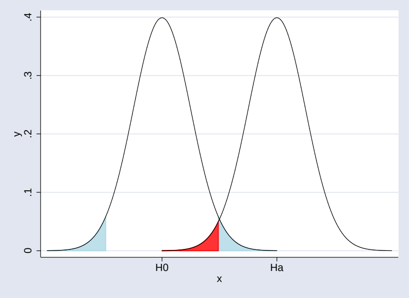
$=$ probability of type I error (If H_0 is true)

Power Calculations Illustrated: 2



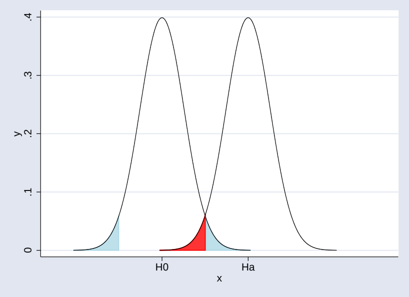
- H_0: $\bar{x} = 0$, S.E.$(x) = 1$
- H_A: $\bar{x} = 3$, S.E.$(x) = 1$
- Power = 85%

Power Calculations Illustrated: 3



- H_0: $\bar{x} = 0$, S.E.$(x) = 1$
- H_A: $\bar{x} = 4$, S.E.$(x) = 1$
- Power = 98%

Power Calculations Illustrated: 4



- H_0: $\bar{x} = 0$, S.E.$(x) = 0.77$
- H_A: $\bar{x} = 3$, S.E.$(x) = 0.77$
- Power = 97%

Power Calculations in Stata

- `sampsi`
- Only for differences between two groups
- Difference in proportion or mean of normally distributed variable
- Can calculate sample size for given power, or power for given sample size
- Does not account for matching
- Need hypothesised proportion or mean & SD in each group

Parameters in Sample Size Calculation

- Power
- Significance level
- Mean (proportion) in each group
- SD in each group
    - Missing SD $\Rightarrow$ proportion

`sampsi` syntax for sample size

`sampsi #1 #2, [ratio() sd1() sd2() power()]`

where

- `ratio` Ratio of the size of group 2 to size of group 1 (defaults to 1).
- `sd1` Standard deviation in group 1 (not given for proportions).
- `sd2` Standard deviation in group 2 (not given for proportions). Assumed equal to `sd1` if not given).
- `power` Desired power as a probability (i.e. 80% power = 0.8). Default is 90%.

`sampsi` for power

`sampsi #1 #2, [ratio() sd1() sd2() n1() n2()]`

where

- `ratio` Ratio of the size of group 2 to size of group 1 (defaults to 1).
- `sd1` Standard deviation in group 1 (not given for proportions).
- `sd2` Standard deviation in group 2 (not given for proportions). Assumed equal to `sd1` if not given).
- `n1` Size of group 1
- `n2` Size of group 2

`sampsi` examples

- Sample size needed to detect a difference between 25% prevalence in unexposed and 50% prevalence in exposed:
  `sampsi 0.25 0.5`
- Sample size needed to detect a difference in mean of 100 in group one and 120 in group 2 if the standard deviation is 20 and group 2 is twice as big as group 1
  `sampsi 100 120, sd1(20) ratio(2)`

Criticism of hypothesis and significance testing

- Hypothesis and significance testing are complicated, convoluted and poorly understood
- Tell us *one* fact that is unlikely to be true about our population
- *p*-value depends on both effect size and study size: contains no explicit information about either
- Intended as automated decision-making process: cannot include other information to inform decision
- Would prefer to know things that are true about the population
- More useful to have a range within which you believe a population parameter lies.

Hypothesis Tests and Confidence Intervals

- Hypothesis tests about means and proportions are closely related to the corresponding confidence intervals for the mean and proportion.
- *p*-value mixes together information about sample size and effect size
- Dichotomising at $p = 0.05$ ignores lots of information.
- Confidence intervals convey more information and are to be preferred.
- There is a movement to remove "archaic" hypothesis tests from epidemiology literature, to be replaced by confidence intervals.
- If there are several groups being compared, a single hypothesis test is possible, several confidence intervals would be required.

Hypothesis Tests vs. Confidence Intervals

Outcome	Exposed	
	No	Yes
No	6	6
Yes	2	5

- p = 0.37
- OR = 2.5, 95% CI = (0.3, 18.3)

Outcome	Exposed	
	No	Yes
No	600	731
Yes	200	269

- p = 0.36
- OR = 1.1, 95% CI = (0.9, 1.4)

The ASA's Statement on *p*-Values: Context, Process, and Purpose

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations

- Survey of 25 most common errors in statistical inference
- 20 involve hypothesis tests, 5 involve confidence intervals
    - E.g. effect significant in men, not significant in women: this is not evidence of a difference in effect between men and women.
    - Same null hypothesis tested many times in different studies, all results are non-significant: not evidence for null hypothesis