Summary of hypothesis tests

Quick reference

One variable

Continuous data

- **One Sampled t-test**: Compare the mean of a sample to a known value or theoretical expectation
- **Paired t-test**: Compare the means of the same group at two different times (e.g., before and after a treatment)
- **One-Sample Wilcoxon test**: Non--parametric test for when the data does not meet the normality assumption. Compare the median of a single column of data to a hypothetical medians

Categorical data

- **Chi-square goodness of fit**: Test whether the observed proportion of categorical data matches an expected proportion
- **Binomial test**: Test whether the probability of success in a binomial experiment is equal to a specific value

Two variables

Continuous - continuous

- **Independent two-sample t-test**: Compare the means of two independent groups
- **Paired t-test**: Compare the means of the same group at two different times (e.g., before and after a treatment)
- **Pearson correlation**: test if two continuous variables are correlated
- **Spearman rank correlation**: Non-parametric test to see if two continuous or ordinal variables are monotonically related
- **Mann-Whitney U test**: Non-parametric alternative to the independent two-sampled t-test

Categorical - Categorical

- **Chi-square test for independence**: Test the independence of two categorical variables
- **Fisher's exact test**: Similar to the Chi-square test but used when sample sizes are small

Categorical - Continuous

- **Independent two-sample t-test**: compare the means of a continuous variable for two categories
- **ANOVA (Analysis of Variance)**: compare the means of a continuous variable for more than two categories
- **Mann-Whitney U test or Kruskal-Wallis Test**: Non-parametric alternatives for the two-sample t-test and ANOVA, respectively

More than two variables

- **ANOVA (Analysis of variance)**: Test if the means of a continuous variable are different for different categories (more than two) of a categorical variable
- **Multiple Regression**: Test the effect of multiple continuous predictors on a continuous outcome
- **Logistic Regression**: Test the effect of multiple continuous or categorical predictors on a binary outcome
- **Multivariate ANOVA (MANOVA)**: An extension of ANOVA that covers situations where there is more than one dependent variable to be tested

Details, examples, assumptions, and caveats

Chi-Square test for independence

- Example: Gather a bunch of robots and a bunch of humans and ask them what they prefer: flowers, puppies, or a properly formatted data file. Do robots and humans have the same preferences?
  - Generating the crosstab gives us this dataset:

	Robot	Human	Total
Puppy	O_{11}	O_{12}	R_1
Flower	O_{21}	O_{22}	R_2
Data file	O_{31}	O_{32}	R_3
Total	C_1	C_2	N

  - O = count of the number of respondants meting that condition
  - C = column totals
  - R = row totals
  - N = sample size
  - Null hypothesis: All of the following statements are true:

        - $P_{\text{robot,puppy}} = P_{\text{human,puppy}}$
        - $P_{\text{robot,flower}} = P_{\text{human,flower}}$
        - $P_{\text{robot,data file}} = P_{\text{human,data file}}$
- test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

- degrees of freedom: $df = (r-1)(c-1)$

Assumptions of the Chi-Squared test for independence

- Expected frequencies are sufficiently large -> normally we want $N > 5$
- observations are independent
- **if the independence assumption is violated** -> try looking into the McNemar test or Cochran test

Chi-Square test for goodness of fit

- Example: Ask people to draw 2 cards from a standard deck at random. Were the cards really drawn at random?
  - Null hypothesis: all four suits are chosen with equal probability (e.g. P = (0.25, 0.25, 0.25, 0.25))
  - Alternative hypothesis: At least one of the suit-choice probabilities ISN'T 0.25
  - Test statistic: Compare expected number of observations in each category (E_i) with the observed number of observations (O_i)
  - Derive the chi-squared statistic using:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Since O_i and E_i represent the probability / frequency of success, they come from a binomial distribution. When $\text{number of samples} \times \text{probability of success}$ is large enough, this becomes a normal distribution. And squaring things that come from normal distributions and adding them up gives you a chi-squared distribution
- degrees of freedom
    - main idea: calculate DoF by counting the number of distinct quantities used to describe the data and subtract off all the constraints.
    - For this case, we describe the data using 4 numbers corresponding to the observed frequencies of each category. There is one fixed constraint: if we know the sample size, we can figure out how many people chose spades given we know how many

people picked hearts, clubs, diamonds, etc. Therefore our degrees of freedom are $N-1$ for N variables plus the constraint that the sum of the probabilities must sum to 1.

- This is always a 1-sided test

Assumptions of the Chi-Squared goodness-of-fit test

- Expected frequencies are sufficiently large -> normally we want $N > 5$
- observations are independent
- **if the independence assumption is violated** -> try looking into the McNemar test or Cochran test

Correction to chi-squared tests when there's 1 DoF

This is called the **Yates Correction** or **continuity correction**. Basically, the chi-squared test is based on the assumption that the binomial distributions look like normal distributions with large N. When there's 1 DoF (i.e, a 2x2 contingency table) and N is small, the test statistic is generally too big. **Yates** proposed this correction as more of a hack, probably not derived from anything and just based on empirical evidence:

$$\chi^2 = \sum_i \frac{j(|E_i - O_i| - 0.5)^2}{E_i}$$

Fisher's exact test

- Use this when you don't have enough samples to do a chi-squared test
- Start with the same contingency table

	Happy	Sad	Total
Set on fire	O_{11}	O_{12}	R_1
Not set on fire	O_{21}	O_{22}	R_2
Total	C_1	C_2	N

- Calculate the probability that we would have obtained the observed frequencies that we did ($O_{11}, O_{12}, O_{21}, O_{22}$) given the row and column totals:

$$P(O_{11}, O_{12}, O_{21}, O_{22} | R_1, R_2, C_1, C_2)$$

- This is describe by a hypergeometric distribution

McNemar test

answers: I'd like to do a chi-squared test but the experiment is repeated measures (e.g., measuring a participant before and after a treatment)

- **Note that this is the same question as a paired-samples t-test with categorical data**
- The trick is to re-label the the data such that each participant (thing we're observing) appears in only one cell

before:

	Before	After	Total
Yes	30	10	40
No	70	90	160
Total	100	100	200

after:

	Before: Yes	Before: No	Total
After: Yes	5	5	10
After: No	25	65	90
Total	30	70	100

label the entries:

	Before: Yes	Before: No	Total
After: Yes	a	b	$a+b$
After: No	c	d	$c+d$
Total	$a+c$	$b+d$	n

Null hypothesis says that the "before" and "after" test have the same proportion of people saying "yes". In other words, the row and column totals come from the same distribution:

$$H_0 : P_a + P_b = P_a + P_c \quad \text{and} \quad P_c + P_d = P_b + P_d$$

This simplifies to:

$$H_0 : P_b = P_c$$

In other words, **We only have to check that the off diagonal entries are equal**

Now this is a normal χ^2 test with the Yates correction:

$$\chi^2 = \frac{(|b-c| - 0.5)^2}{b+c}$$

Z-test

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$$

- Assumes that you know the population standard deviation
- *What if you don't know the population standard deviation like in 99.9% of experiments? ->* **Use a T-test**

One-sample t-test

- **asks: Does this sample have this population mean?**

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

- this is a t-distribution with $N - 1$ degrees of freedom.

Assumptions

- *Normality* - assume that the population distribution is normal
- *Independence* - observations are independent of each other

Independent samples t-test (Student test)

- **asks: Do these two groups have the same population mean?**
- Hypotheses:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}}$$

Figuring out the standard error can be a bit tricky:

$$\mathrm{SE}(\bar{X}_1 - \bar{X}_2) = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

where

$$\hat{\sigma}^2 = \frac{\sum_{ik}(X_{ik} - \bar{X}k)^2}{N - 2}$$

- Note: having a negative value for the t statistic isn't a big deal, it just means that the group mean for X_2 is bigger than X_1.

Assumptions

- *Normality* - assume both groups are normally distributed

- *Independence* - assume observations are independently sampled (including no cross sampled observations (e.g. participants in both group 1 and 2))
- *Homogeneity of variance / homosccedasticity*: population standard deviations are the same for both groups (test this using Levene test)

Independent samples t-test (Welch test)

- **asks: Do these two groups have the same population mean? (assuming they have different variances)**
- What if your two groups don't have equal variances? (e.g. violate the 3rd assumption of the previous test)
- Use the same t-statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}(\bar{X}_1 - \bar{X}_2)}$$

where:

$$\mathrm{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

degrees of freedom:

$$df = \frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{(\hat{\sigma}_1^2/N_1)^2/(N-1) + (\hat{\sigma}_2^2/N_2)^2/(N_2-1)}$$

Assumptions

- *Normality* - assume both groups are normally distributed
- *Independence* - assume observations are independently sampled (including no cross sampled observations (e.g. participants in both group 1 and 2))

Paired-samples t-test

asks: are the means from a repeated measures design the same? (e.g., measure a person at time x, apply a treatment, measure a person at time y, do the populations at times x and y have the same mean?)

- Run a one-sampled t test with a difference variable, called *improvement*

$$D_i = X_{i1} - X_{i2}$$

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

$$t = \frac{\bar{D}}{\mathrm{SE}(\bar{D})} = \frac{\bar{D}}{\hat{\sigma}_D/\sqrt{N}}$$

Assumption tests

Levene test

asks: Do my groups have the same population variance?

Test statistic:

$$Z_{ik} = |Y_{ik} - \bar{Y}_k|$$

Z_{ik} is a measure of how the i-th observation in the k-th group deviates from its group mean, so:

$$\begin{aligned} H_0 &: \text{The population means of Z are identical for all groups} \\ H_1 &: \text{The population means of Z are NOT identical for all groups} \end{aligned}$$

Note that the test statistic is calcuated in the same way as the F-statistic for regular ANOVA

- this is a good check to run if you're not sure that the data is normal ref

Brown-Forsythe test

asks: Do my groups have the same population variance?

Same as the levene test, but uses the group median:

$$Z_{ik} = |Y_{ik} - \mathrm{median}_k(Y)|$$

Fmax test (aka Hartley's test)

asks: Do my groups have the same population variance?

The hypotheses are:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \ldots \\ H_1 &: \text{population variances are not equal} \end{aligned}$$

test statistic (follows an F distribution)

$$F_{\max} = \frac{\sigma^2_{\max}}{\sigma^2_{\min}}$$

with $df = N - 1$

assumptions

- number of samples drawn from each population is roughly the same
- populations are normally distributed (very sensitive to this)

Bartlett test

- this is a good check to run if you're not sure that the data is normal [ref]

References

- Multiple conversations with chatGPT
- Learning Statistics with R, chapter 12.1, 12.7, 13
- https://accendoreliability.com/hartleys-test-variance-homogeneity/