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:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

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

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*

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

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

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

Two-Sample Wilcoxon Test (aka Mann-Whitney test)

Asks: I'd like to run a t-test, but the data does not follow a normal distribution

This is a non-parametric alternative to a t-test

Suppose we have 2 groups and some scores:

score	group
1	A
2	B
2.5	A
2.2	B

All we have to do a make a table that compares every observation in each group, and mark where group A is greater.

	2 (group B)	2.2 (group B)
1 (group A)		
2.5 (Group A)	x	x

There are 2 marks, so our test statistic is 2.

The actual sampling is complicated, just plug this into a computer and it'll get you the p-value.

One-sample Wilcoxon test (aka paired samples Wilcoxon test)

Asks: I want to do a paired-samples t-test, but the data does not follow a normal distrubtion

Construct this the same way you would construct a 2-sample wilcoxon test, where one variable is the positive differences between the two measurements and the other variable is all the differences

One-way ANOVA

Asks: I have several groups of observations, do these groups differ in terms of some outcome variable of interest?

see One-way ANOVA a more detailed explaination

We can attribute the variance observed when fitting the model to two sources:

- 1. differences in the the groups (**between group SS**)
- 2. differences within each group (**within-group SS**)

If the differences between the groups constitutes a large enough proportion of the total variation, then we can conclude that the difference is actually due to different means

- This is similar to linear regression but we're using a categorical variable to predict a continuous variable.

Assumptions

- Residuals are normally distributed (use QQ plot or Shapiro-Wilk test)
- homogeneity of variance / homoscedasicity
- independence of residuals

Kruskal-Wallis test

Asks: I want to do ANOVA, but the data is not normally distributed

This is a non-parametric test to compare the means of 3 or more groups

- In ANOVA, we started with Y_{ik} - the outcome of the ith person in the kth group. Now, rank order those values and do the analysis on the ranked data.
- Let R_{ik} be the ranking of the ith member in the kth group. Calculate the average rank given to the observations in the kth group:

$$\bar{R}_k = \frac{1}{N_k} \sum_i R_{ik}$$

also calculate the grand mean rank

$$\bar{R} = \frac{1}{N} \sum_i \sum_k R_{ik}$$

Now we can calculate variances from the mean rank:

How far the ikth observation deviates from the grand mean rank

$$\mathrm{RSS}_{\mathrm{tot}} = \sum_k \sum_i (R_{ik} - \bar{R})^2$$

How much the group deviates from the grand mean rank

$$\mathrm{RSS_b} = \sum_k \sum_i (\bar{R}_k - \bar{R})^2 = \sum_k N_k (\bar{R}_k - \bar{R})^2$$

Now we'll build the test statistic as a comparison between how much the group deviates from the grand mean rank vs. how much the individual samples deviate

$$K = (N-1) \times \frac{\mathrm{RSS_b}}{\mathrm{RSS_{tot}}}$$

This means that if K is sufficiently large, then the group deviations explains a lot of the variance, so we may conclude that the differences are real.

K follows approximately a χ^2 distribution with $G-1$ degrees of freedom (G = number of groups)

Sometimes you'll see K written as (after some algebraic magic):

$$K = \frac{12}{N(N-1)} \sum_k N_k \bar{R}_k^2 - 3(N+1)$$

But what if there are ties?
Compute a frequency table with the observed value and number of times you observed it, call this f_j, so if you observe the value 4 three times, then f_4 = 3

Compute the tie correction factor (TCF):

$$\mathrm{TCF} = 1 - \frac{\sum_j f_j^3 - f_j}{N^3 - N}$$

and divide K by this value:

$$K_{\mathrm{corrrected}} = \frac{K}{\mathrm{TCF}}$$

Pearson Correlation

Asks: What is the relationship between the two variables, and is it significant?

$$r_{\mathrm{XY}} = \frac{\mathrm{Cov}(X,Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

$r=1$ is a perfectly positive relationship, $r=-1$ is a perfectly negative relationship, and $r=0$ is no relationship

If you run a hypothesis test for the significance of this correlation, **it is identical to the t-test that's run on a single coefficient of a regression model**

Spearman's rank correlation

Asks: What is the relationship between two variables, and how can we tell whether the relationship is ACTUALLY linear?

To elaborate on the asks section, think about the 80/20 rule. 20% effort can lead to 80% improvement - this is NOT a linear relationship. Another way to think of it is we want to capture the idea of diminishing returns.

As another example, consider **Anscombe's quartet** - it's a set of 4 completely different datasets that all have a pearson correlation of $r = 0.816$

To account for this, instead of comparing X with Y, compare the RANK of X and Y. For example, if our dataset is:

	hours	grade
1	2	13
2	76	91
3	40	79
4	6	14
5	16	21
6	28	74
7	27	47
8	59	85
9	46	84
10	68	88

The rank data would be:

	rank (hours worked)	rank (grade received)
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

(datasets come from section 5.7.6 of learning statistics with R)

Now we can just do regular person's correlation on this transformed data.

R^2 Value

Asks: How good is my regression model at fitting the data?

Calculate the sum of the squared residuals:

$$\text{SS}_{\text{res}} = \sum_i (Y_i - \hat{Y}_i)^2$$

and the total variability in the outcome variable

$$\text{SS}_{\text{tot}} = \sum_i (Y_i - \bar{Y})^2$$

The **Coefficient of determination**, or R^2, quantifies the proportion of the variance in the outcome variable that can be accounted for by the predictor

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

This is the square of the pearson correlation: $R^2 = \text{cor}(X, y)^2$

Adjusted R^2 Value

Adding more predictors will always cause R^2 to increase, so there's another version of R^2 that takes degrees of freedom into account. If you have K predictors and N observations:

$$\text{adj. } R^2 = 1 - (\frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \times \frac{N-1}{N-K-1})$$

this will only increase if the new variables improve model performance than you'd expect by chance, however you can't interpret it the same way as R^2.

R^2 or adjusted R^2?

Do you want the results to be interpretable? R^2
Do you want to correct for bias? adjusted R^2

Hypothesis test for a regression model

Asks: I just fit a regression model: Y = mx + b. Is there an actual relationship between the predictors and outcome?

Our hypotheses:

$$\begin{aligned} H_0 : Y_i =& b_0 + \epsilon_i \\ H_1 : Y_i =& (\sum_{k=1}^{K} b_k X_{ik}) + b_0 + \epsilon_i \end{aligned}$$

We're going to treat this JUST LIKE ANOVA, except instead of comparing between- and within-group variances, we're going to compare the residual variance and the model variance.

$$\text{SS}_{\text{mod}} = \text{SS}_{\text{tot}} - \text{SS}_{\text{res}}$$

where:

$$\mathrm{SS}_{\mathrm{tot}} = \sum_i (Y_i - \bar{Y})^2$$

and

$$\mathrm{SS}_{\mathrm{res}} = \sum_i (Y_i - \hat{Y}_i)^2$$

now divide by degrees of freedom: $df_{\mathrm{mod}} = K$, $df_{\mathrm{res}} = N - K - 1$

$$\mathrm{MS}_{\mathrm{mod}} = \frac{\mathrm{SS}_{\mathrm{mod}}}{\mathrm{df}_{\mathrm{mod}}}$$

$$\mathrm{MS}_{\mathrm{res}} = \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{df}_{\mathrm{res}}}$$

Finally run the F-test with the test statistic:

$$F = \frac{\mathrm{MS}_{\mathrm{mod}}}{\mathrm{MS}_{\mathrm{res}}}$$

Hypothesis test for regression coefficients

Asks: Is this specific coefficient in a linear regression model meaningful?

Example: let's say we fit the model: $y = 5x_1 + 0.1x_2 + 3$. Is there actually a relationship between x_2 and y?

Our null hypothesis is that the true regression coefficient is 0:

$$\begin{aligned} H_0 &: \quad b = 0 \\ H_1 &: \quad b \neq 0 \end{aligned}$$

If we assume that the sampling distribution of coefficient is normal (which is a safe assumption given the central limit theorem), then **This is just a t-test**

$$t = \frac{\hat{b}}{\mathrm{SE}(\hat{b})}$$

degrees of freedom = $df = N - K - 1$

Standard error of the regression coefficient is complex to calculate. Based on the footnote in learning statistics with R it goes something like this:

- The vector of residuals is $\epsilon = y - X\hat{b}$
- Residual variance is: $\hat{\sigma}^2 = \epsilon^T \epsilon / (N - K - 1)$
- Covariance matrix for the coefficients is: $\hat{\sigma}^2 (X^T X)^{-1}$
- The diagonal of these is $\mathrm{SE}(\hat{b})$ - our standard error

- **This reminds me of the formula for PCA - maybe there's a connection?**

This test is identical to the test for the significance of a pearson correlation

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:

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

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

The steps for this test are pretty extensive, but it's another test for equality of variance and is a good option if you're not sure that the data is normally distributed.

See the steps here:
https://accendoreliability.com/bartletts-test-homogeneity-variances/

Post-hoc tests

Asks: I ran anova with more than two groups and got a significant result. *which groups are actually different?*

The simplest approach is to run a series of t-tests between pairs of groups. So if you have groups A, B, and C - you'll want to do t-tests comparing groups A & B, A & C, B & C, etc.

A word of caution: when you start running post-hoc tests you'll start running lots and lots of tests (like a fishing expedition) without a lot of theoretical guidance. If you run 45 t-tests to check an anova from 10 different predictors, you'd expect 2-3 of them to be significant by chance alone. When this happens, *Your type-I error rate has gotten out of control*

Bonferroni's correction

Adjust your p-values to account for multiple tests. If you run m separate tests and the original p-value is p, then your adjusted p - value is:

$$p' = m \times p$$

and reject the null hypothesis if $p' < \alpha$

Holm correction

This is normally used more frequently than Bonferroni. It has the same type I error rate but it has a lower type II error rate (more *powerful* than Bonferroni)

To perform the correction:

- Sort all the p-values in order from smallest to largest
- For the smallest p-value, multiply it by m (the number of tests you ran) - then you're done
- For the larger p-values:
  - multiply the p-value by $m - n$, where n is the number of times you've iterated so far.

    - If this number is bigger than the adjusted p-value from last time, keep it.
    - If this number is smaller, then copy the last p-value
    - repeat

Turkey's HSD (Honestly Significant Difference) test

Asks: Are the pairwise differences between two groups significantly different?

- Constructs simultaneous confidence intervals for all comparisons
- 95% simultaneous confidence interval means that there's a 95% probability that ALL of the confidence intervals contain the relevant true value.
- can use these to calculate an adjusted p value for any comparison

References

- Multiple conversations with chatGPT
- Learning Statistics with R, chapters 5, 12.1, 12.7, 13, 14, 15
- https://accendoreliability.com/hartleys-test-variance-homogeneity/
- https://www.youtube.com/watch?v=oOuu8IBd-yo