Faculty of Health Sciences

Day 2: Hypothesis testing, tests for continuous responses, multiple testing

Paul Blanche

Section of Biostatistics, University of Copenhagen

October 25, 2023

Outline/Intended Learning Outcome (ILOs)

Hypothesis testing
ILO: to describe the principles and logic of hypothesis testing

One and two sample tests for continuous responses: t-test
ILO: to identify when, how and why to use a t-test
ILO: to define a p-value and contrast its use with that of a confidence interval

Power and Sample size calculation
ILO: to identify why and how to make power and sample size calculations
ILO: to analyse their strengths and limitations

Multiple testing
ILO: to describe the multiple testing problem and employ basic remedies

Nonparametric test: Wilcoxon
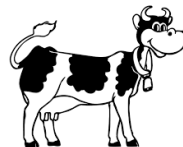ILO: to contrast pros and cons of Wilcoxon vs t-test

Case: cow milk data

▶ **Research question:**
Should cows be fed with Barley or Lupin, to
produce the best milk?

▶ **Outcome:**
protein level of the milk (%) at 12 weeks after
calving.

Statistical aim: provide a yes/no answer about the population
supported by the observed data (sample) while controlling the risks of a
"false finding", via a Hypothesis test.[1]

---

[1]**Note:** important complementary information is given by the confidence interval of
the effect size.

Research question and Null hypothesis

▶ A hypothesis test aims to answer a very precise & specific research question.

   Case: Is there a **difference in (population) mean** level of protein between cows fed with lupin and barley, at 12 weeks?

▶ The null hypothesis $\mathcal{H}_0$ of the test should reflect it and state the opposite of what you aim to prove.

   ▶ Scientific hypothesis: there is a difference.

   ▶ Null hypothesis: there is **no** difference.

   Choosing the opposite is important to appropriately control the **risk of wrong conclusion.**

Hypothesis testing and risks of false conclusions
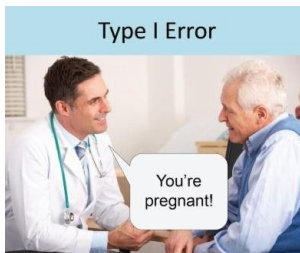
Case:

► Type-I error: conclude a difference although it does not exist.

  → False positive finding

► Type-II error: do not conclude to a difference although it exists.

  → False negative finding

Hypothesis testing and risk control

We want to ensure that the risk of wrongly rejecting the null hypothesis (α) is small (often 5%), i.e. a small risk of a false scientific finding.

Reasoning: the data need to be convincing enough to support the (new) research finding.

Limitation: it might be difficult to have enough data to support a (new) finding ($\rightarrow$ power).

The logic of hypothesis testing

1. **Assume** that the data have been generated in a world in which the **null hypothesis is true**.

2. Under this assumption, **calculate how unlikely** it is to obtain some results that **contradict the null hypothesis** as least as much as those obtained with your data (i.e., compute the p-value).

3. Reject the null hypothesis if this is unlikely 'enough'.

▶ Similar to a proof by contradiction.
▶ Computation in step 2. depends on the type of observed data.

Outline/Intended Learning Outcome (ILOs)

Hypothesis testing
    ILO: to describe the principles and logic of hypothesis testing

One and two sample tests for continuous responses: t-test
    ILO: to identify when, how and why to use a t-test
    ILO: to define a p-value and contrast its use with that of a confidence interval

Power and Sample size calculation
    ILO: to identify why and how to make power and sample size calculations
    ILO: to analyse their strengths and limitations

Multiple testing
    ILO: to describe the multiple testing problem and employ basic remedies

Nonparametric test: Wilcoxon
    ILO: to contrast pros and cons of Wilcoxon vs t-test

Case: cow milk data

Data from $n = 25$ (Barley)+27 (Lupin) cows:

```
protein   Diet
   3.28 lupins
   3.04 barley
   3.07 barley
   2.92 barley
   3.29 lupins
   3.18 lupins
```

etc...

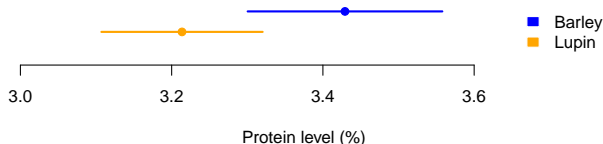|	Barley	Lupin
Mean (SD):	3.43 (0.31)	3.21 (0.27)

▶ Is the difference observed in the data sample large enough to conclude to a difference in the population?

First approach (not optimal for testing)

Comparison of 95% confidence intervals:

- Lupin: [3.11;3.32]
- Barley: [3.30;3.56]

We cannot conclude on the significance of the difference

(see slides lecture 1).

But the two CI can be interesting to report anyway.

A better approach

Compute:

- ▶ p-value for the difference in mean.
- ▶ confidence interval for the difference in mean.

Two-sample t-test (1/2)

Model assumptions: (1 & 2 are important, 3 not always)

1. The two samples are independent (no pairing).
2. Observations from each sample are independent.
3. Observations are normally distributed.

To test the null hypothesis $\boxed{\mathcal{H}_0 : \mu_1 = \mu_2}$, i.e. the population means are the same in the two populations, we compute the t-statistic.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s.e.(\bar{x}_1 - \bar{x}_2)}$$

where the standard error is $s.e.(\bar{x}_1 - \bar{x}_2) = \sqrt{s_1^2/n_1 + s_2^2/n_2}$.

The value t quantifies how large the (sample) difference $(\bar{x}_1 - \bar{x}_2)$ is relative to the amount of information provided by the data $(s.e.(\bar{x}_1 - \bar{x}_2))$ and is used to compute a p-value.

Two-sample t-test (2/2)

The key idea to use the t-statistics is that under the model assumption, it follows a specific distribution[2] whatever the value of the (population) means (μ_1, μ_2) and standard deviations (σ_1, σ_2) in each group.

Hence we can assume $\mu_1 = \mu_2$ and calculate how unlikely it is to obtain a t value that contradicts the null hypothesis as least as much as that obtained with your data, that is, we can compute a **p-value**.

The larger $|t|$ the more the data contradict $\mathcal{H}_0 : \mu_1 = \mu_2$.

p-value= $P(|T| > |t|)$, where T is a random variable that follows the t-distribution.

---

[2] the t-distribution aka Student's distribution, which depends on the two sample sizes n_1 and n_2; already encountered in Lecture 1.

The p-value (1/2)

Informal definition:
The p-value is *"the probability of seeing a result as extreme as your observed result, when the null model is true."* [3]

- *"result"*= test statistic, i.e., a single value that ("cleverly") summarizes the data.
- *"extreme"*= unlikely/unexpected.

Interpretation:
1. If the **p-value is small** the data are at odds with the null hypothesis and the finding is said to be statistically significant.
2. If the **p-value is large**, the finding is said to be not statistically significant.

3 Westfall & Henning. Understanding advanced statistical methods. CRC Press, 2013 (page 405).

The p-value (2/2)

Interpretation:

3. We imagine a large number of repetitions of the study with the null hypothesis being true and define the p-value as the proportion of these studies which provide less support for the null hypothesis than the data actually observed.

Recommendations:

▶ Traditionally the value p=5% has been used to divide "significant" from "non-significant" results, but good practice is to report the actual p-value.

▶ The choice of the threshold to claim significance (e.g., p=5%) should be prespecified.

p-value and strength of evidence

The smaller the p-value, the stronger the evidence against the null hypothesis.

Figure 8.2 in Kirkwood & Sterne (2003), *Essential medical statistics, 2nd edition.*
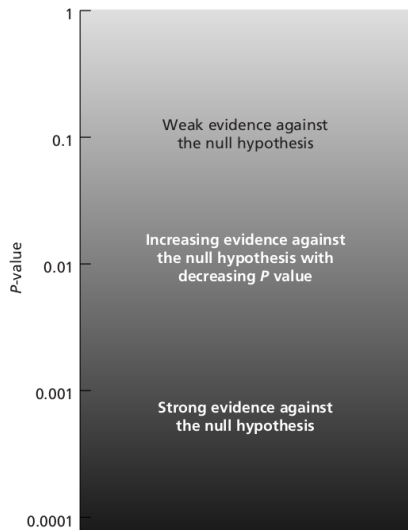
Case: Two-sample t-test

- $\bar{x}_1 = 3.43,\ \bar{x}_2 = 3.21$
- $\bar{x}_1 - \bar{x}_2 = 0.22$
- $n_1 = 25,\ n_2 = 27$
- $s_1 = 0.31,\ s_2 = 0.27$
- $s.e.(\bar{x}_1 - \bar{x}_2) = 0.081$
- $t = 2.66$
- p-value= $P(|T| > |t|)$=0.011

We conclude that there is a significant difference in mean protein level of the milk between cows fed with barley and lupin (p=0.011).

Don't misinterpret p-values!

Widespread misunderstandings called for clarifications:

▶ The p-value is not the probability of $\mathcal{H}_0$ being true.

  "Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither."[4]

  This is because $P(\text{extreme result}|\mathcal{H}_0) \neq P(\mathcal{H}_0|\text{extreme result})$.

18/54
[4] Wasserstein & Lazar. "The ASA statement on p-values: context, process, and purpose." Am. Stat. 70.2 (2016): 129-133.
[5] Pawel & Schwab (2020), Significance 17(6), 10-11; Goodman (1992) Statistics in Medicine, 11(7), 875–879.

Don't misinterpret p-values!

Widespread misunderstandings called for clarifications:

▶ The p-value is not the probability of $\mathcal{H}_0$ being true.

"Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither."[4]

This is because $P(\text{extreme result}|\mathcal{H}_0) \neq P(\mathcal{H}_0|\text{extreme result})$.

▶ Observing a p-value $\leq 5\%$ does not mean that the positive finding is easy to reproduce! [5]

As an example, suppose you conduct a study and find a significant p-value of $p = 5\%$. You try to replicate your positive finding by conducting the exact same study again. What is your chance to get a significant p-value ($\leq 5\%$) again? To calculate that, you need to know the true treatment effect "$\mu_1 - \mu_2$", which you do not know, of course. But what about if we assume that it is what you have estimated in your (first) study, i.e. $\bar{x}_1 - \bar{x}_0$?

[4] Wasserstein & Lazar. "The ASA statement on p-values: context, process, and purpose." Am. Stat. 70.2 (2016): 129-133.
[5] Pawel & Schwab (2020), Significance 17(6), 10-11; Goodman (1992) Statistics in Medicine, 11(7), 875–879.

Don't misinterpret p-values!

Widespread misunderstandings called for clarifications:

▶ The p-value is not the probability of $\mathcal{H}_0$ being true.

   "Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither."[4]

   This is because $P(\text{extreme result}|\mathcal{H}_0) \neq P(\mathcal{H}_0|\text{extreme result})$.

▶ Observing a p-value $\leq 5\%$ does not mean that the positive finding is easy to reproduce! [5]

   As an example, suppose you conduct a study and find a significant p-value of $p = 5\%$. You try to replicate your positive finding by conducting the exact same study again. What is your chance to get a significant p-value ($\leq 5\%$) again? To calculate that, you need to know the true treatment effect "$\mu_1 - \mu_2$", which you do not know, of course. But what about if we assume that it is what you have estimated in your (first) study, i.e. $\bar{x}_1 - \bar{x}_0$? Only 50%!
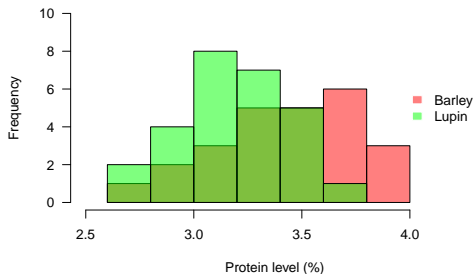
---

[4] Wasserstein & Lazar. "The ASA statement on p-values: context, process, and purpose." Am. Stat. 70.2 (2016): 129-133.
[5] Pawel & Schwab (2020), Significance 17(6), 10-11; Goodman (1992) Statistics in Medicine, 11(7), 875–879.

Normality assumption

Normality should be checked for each sample separately (using histograms or qqplots).

But, when sample sizes n_1 and n_2 are both large enough (say > 15) normality is **not important**[6].

However, **skewed data can be transformed** to facilitate the interpretation and reduce the influence of outliers (this should be pre-specified).

[6]due to the central limit theorem.

Confidence interval of the difference

Good practice: report an estimate of the mean difference and a 95% confidence interval.

$$\bar{x}_1 - \bar{x}_2 \pm t_{df} \cdot s.e.(\bar{x}_1 - \bar{x}_2)$$

▶ df: degree of freedom $\approx n_1 + n_2 - 2$ when $n_1 = n_2$ and $s_1 = s_2$.

▶ $t_{df} \approx 1.96$ when n_1 and n_2 are large (say ≥ 15).

▶ software will take care.

Case: mean difference of -0.22 (CI-95% = [0.05;0.38]; p-value = 0.011).

Confidence interval vs p-value

▶ if 0 is $\left\{ \begin{array}{c} \textbf{in} \\ \textbf{not in} \end{array} \right\}$ the CI, then the difference $\left\{ \begin{array}{c} \textbf{is not} \\ \textbf{is} \end{array} \right\}$ significant.

▶ We can tell if the test is significant from looking at the CI, but we can't guess the CI from knowing the p-value.

Confidence interval vs p-value

▶ if 0 is $\left\{ \begin{array}{c} \textbf{in} \\ \textbf{not in} \end{array} \right\}$ the CI, then the difference $\left\{ \begin{array}{c} \textbf{is not} \\ \textbf{is} \end{array} \right\}$ significant.

▶ We can tell if the test is significant from looking at the CI, but we can't guess the CI from knowing the p-value.

▶ A wide 95% that includes 0 suggests "lack/absence of evidence".

▶ A narrow 95% that includes 0 suggests "evidence of absence" of difference (or existence of a "tiny one", if any).

Difference in mean
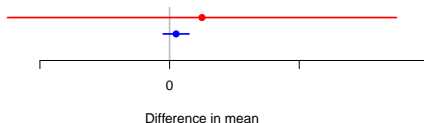
Altman & Bland (1995). *Statistics notes: Absence of evidence is not evidence of absence.* BMJ, 311(7003), 485.

Two versions of the two-sample t-test

"Classical" Student's t-test (not recommended):

▶ Original t-test, described in many basic textbooks.

▶ Additional assumption of equal standard deviations $\sigma_1 = \sigma_2$.

▶ Different formula for s.e. and degrees of freedom ($df = n_1 + n_2 - 2$).

Welch's t-test (the presented one, recommended):

▶ No assumption of equal standard deviations: less restrictive.

▶ Formula for degrees of freedom more complicated, but software take care.

▶ Default in R.

One-sample example

Research question:
Is the mean protein level of the milk similar at 1 and 12 weeks after calving, for cows fed with Barley?

Data $(t_1 - t_{12})$:

```
Cow  Diff
B01 -0.08
B02 -0.03
B03  1.06
B04  0.48
B05  0.49
B06  0.74

etc...
```

Null hypothesis:
The mean difference between protein level at 1 and 12 weeks is zero ($\mathcal{H}_0 : \mu = 0$).

One-sample test because only one group of ($n =25$) cows (barley).

One-sample t-test

The t-test statistic measures the distance between the sample mean and
the assumed population mean μ under $\mathcal{H}_0$ in units of the standard error:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

If $|t|$ is large, the data "contradict" the null hypothesis.

$$\text{p-value} = P(|T| > |t|)$$

where T is a random variable that follows the t-distribution with $n - 1$
degrees of freedom.

▶ similar to the computation of the confidence intervals for the mean.
▶ p-value $\leq 5\%$ $\iff$ μ not in 95% CI.

One-sample t-test: example results

- $\bar{x} = 0.46$
- $n = 25$
- $s = 0.31$
- $t = 7.43$
- p-value= $P(|T| > |t|) < 0.001$ (for $\mathcal{H}_0 : \mu = 0$)

We conclude that there is a significant difference in mean protein level of the milk at 1 and 12 weeks after calving, for cows fed with barley (p<0.001).

Reminder (see Lecture 1):
we compute the 95% CI as $\bar{x} \pm t_{n-1} \cdot s/\sqrt{n}$, which here leads to [0.33;0.58] (and does not inlclude 0).

Note: this one-sample t-test corresponds to a **paired t-test**[7].

[7]two samples of observations (two times) paired by cow. More on Lecture 8.

Outline/Intended Learning Outcome (ILOs)

Hypothesis testing
    ILO: to describe the principles and logic of hypothesis testing

One and two sample tests for continuous responses: t-test
    ILO: to identify when, how and why to use a t-test
    ILO: to define a p-value and contrast its use with that of a confidence interval

Power and Sample size calculation
    ILO: to identify why and how to make power and sample size calculations
    ILO: to analyse their strengths and limitations

Multiple testing
    ILO: to describe the multiple testing problem and employ basic remedies

Nonparametric test: Wilcoxon
    ILO: to contrast pros and cons of Wilcoxon vs t-test

Power

The **power** of a test is the chance of obtaining a significant result when the null hypothesis is indeed false.

Power

The **power** of a test is the chance of obtaining a significant result when the null hypothesis is indeed false.

▶ Power $= 1 - \beta$, i.e. 1 minus the risk of a "false negative" result (β), i.e. 1 minus risk of Type-II error.

▶ Although we can control the type-I error ($\alpha = 5\%$) by appropriately computing the p-value and comparing it to 5%, the computation does not control the risk of type-II error, β.

▶ The power of a two-sample t-test depends on:
  ▶ sample sizes n_1 and n_2 (the larger the better).
  ▶ standard deviations σ_1 and σ_2 (i.e. variability, the smaller the better).
  ▶ difference in mean $\delta = |\mu_1 - \mu_2|$ (i.e. effect size, the larger the better).

Textbook power formula (approximation for two-sample t-test)

$$\delta = (z_{1-\beta} - z_{\alpha/2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

▶ $z_{\alpha/2} = -1.96$ for $\alpha = 5\%$.[8]

▶ $z_{1-\beta} = 0.84$ and 1.28 for $1 - \beta = 80\%$ and 90%.

▶ maximal power when $n_1 = n_2$, for a given total sample size $n_1 + n_2$, when $\sigma_1 = \sigma_2$.

Useful for computing:

▶ Sample size: $n_1 = n_2$ for given "guesses" of σ_1, σ_2 and δ and desired $1 - \beta$ and α.

▶ Power for a given budget/sample size: $1 - \beta$ for "guesses" of σ_1, σ_2 and δ and desired n_1, n_2 and α.

▶ Least detectable difference: δ for given n_1 and n_2, "guesses" of σ_1 and σ_2 and desired α and minimal power $1 - \beta$.

[8]where z_γ is the γ-quantile of a standard normal distribution.

Use a software ! (e.g. R)

Often it is "good enough" to assume $\sigma_1 = \sigma_2$ and then sensible to choose $n_1 = n_2$. Then standard software can be used, e.g. with R[9]:

```
power.t.test(power = .80, delta = 0.5)

     Two-sample t test power calculation

              n = 63.76576
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

▶ $n_1 = n_2 = 64$ subjects needed to detect $1/2$ sd difference[10].

---

[9] slightly more precise calculation performed than using the textbook formula.

[10] Note: it holds for whatever $\sigma_1 = \sigma_2$ and δ, as long as δ/σ_1, the "signal-to-noise ratio", is $1/2$.

Sample size calculation: which difference δ to use?

Principled choices:

- expected/hypothesized difference.
- minimum (clinically) relevant difference.

But small difference are difficult to detect and may require a large sample size, with consequences on the budget, study length, etc.

Pragmatic choice: smallest difference "disappointing" to overlook.

If this still indicates a too large sample size, then discuss with your supervisor (try to avoid wasting time/money).

Which guesses for the standard deviations?

For the calculations, we need a "guess" for the variability in the outcome[11], i.e. σ_1, σ_2.

▶ Estimate from previous studies from your research group or published in the literature (be aware of statistical uncertainty).

▶ Expert guess (supervisor/senior collaborators).

Recommended practice:

▶ use several likely values to do several calculations.

▶ see how changes affect the results and discuss with your collaborators.

▶ be conservative (when appropriate).

▶ consider ethical issues (when appropriate).

[11] Thinking about the normal range width (=4σ) can help to guess σ.

Least detectable difference: sensitivity to σ

Example: my grant (money/time) can finance a sample size of $n = 40$ (i.e. 20 per group), what is the smallest difference I can hope to show with a decent power (e.g. 80%)?

```
power.t.test(n=20,sd=1,power=0.80)

     Two-sample t test power calculation

               n = 20
           delta = 0.9091306
              sd = 1
       sig.level = 0.05
           power = 0.8
     alternative = two.sided

NOTE: n is number in *each* group
```
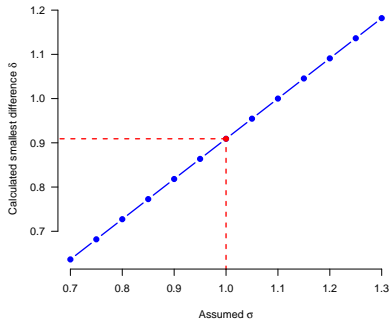

Note: textbook formula gives $\delta = 2.8 \cdot \sigma \cdot \sqrt{2/20}$.

Power: sensitivity to σ

Example: an initial calculation suggests $n = 74$ (i.e. 37 per group), for the minimum difference $\delta = 2$ that we aim to show, with our best expert guess $\sigma = 3$ (with 80% power). But what does the power become if we over or underestimate σ by up to 50%?
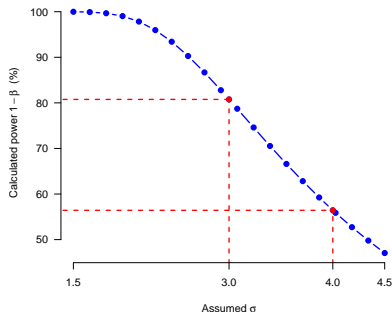
```
power.t.test(sd=4,delta=2,n=37)

    Two-sample t test power calculation

              n = 37
          delta = 2
             sd = 4
      sig.level = 0.05
          power = 0.5642987
    alternative = two.sided

NOTE: n is number in *each* group
```


Note: textbook formula gives $z_{1-\beta} = (2/\sigma) \cdot (\sqrt{37}/\sqrt{2}) - 1.96$ and tables and software give $z_{1-\beta} = 1.64, 1.28, 0.84,$ 0.25, -0.52 for $1 - \beta = 95, 90, 80, 60$ and 30%, respectively.

Outline/Intended Learning Outcome (ILOs)

Hypothesis testing
     ILO: to describe the principles and logic of hypothesis testing

One and two sample tests for continuous responses: t-test
     ILO: to identify when, how and why to use a t-test
     ILO: to define a p-value and contrast its use with that of a confidence interval

Power and Sample size calculation
     ILO: to identify why and how to make power and sample size calculations
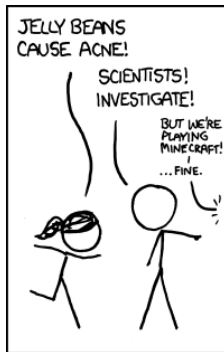   ILO: to analyse their strengths and limitations

Multiple testing
     ILO: to describe the multiple testing problem and employ basic remedies

Nonparametric test: Wilcoxon
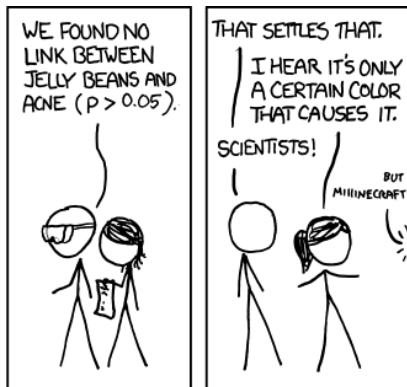   ILO: to contrast pros and cons of Wilcoxon vs t-test
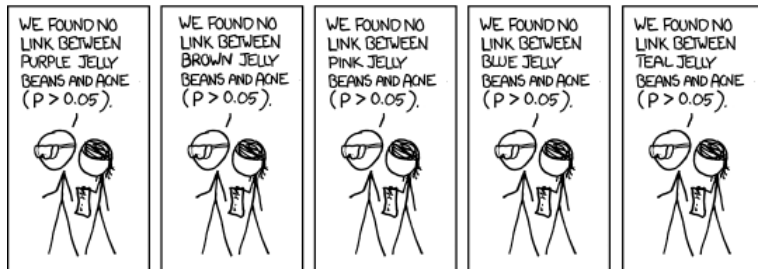
A multiple testing example
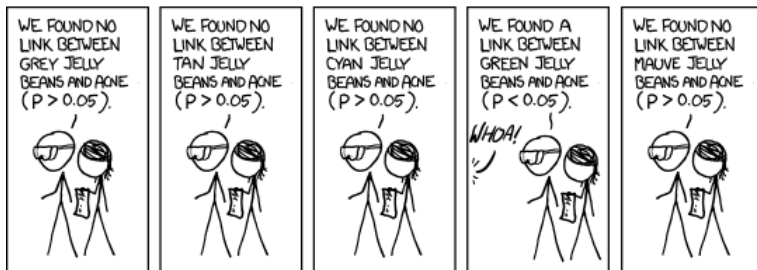

Are jelly beans associated
with acne?

(cartoon from: https://xkcd.com/882/)

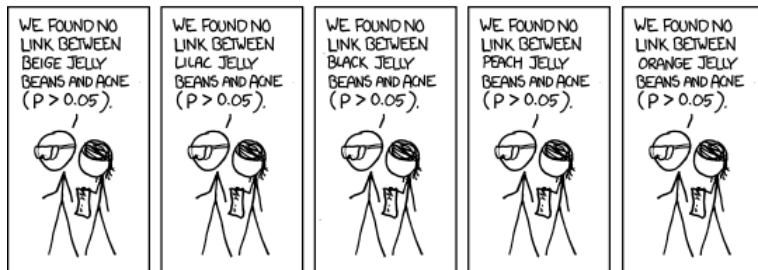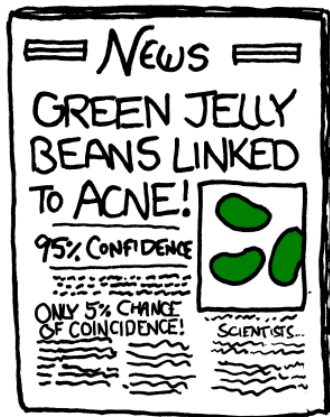- First test is not significant.
- Move on to other tests.

- Five more tests are not significant.
- Move on to other tests.

▶ Four more tests are not significant, but one is significant (Green!).
▶ Move on to other tests.

- Five more tests are not significant.
- Stop testing.

▶ Conclude.

Is the conclusion correct? Why?

Multiple testing issue

▶ The risk of type-I error of **each** test is controlled (usually at 5%).

▶ i.e. thinking of each hypothesis test separately, each corresponding to a specific research question and specific study, the risk of false positive finding is controlled for each of them.

▶ But, if we consider them part of the same study and consider that we have a finding if at least one test is significant, then we do not control the risk of false positive finding.

▶ i.e. the risk of having **at least one** significant p-value although there is no association is not controlled.

Family-wise error rate (FWER): probability of making one or more false discoveries when performing multiple hypotheses tests

FWER in the example

We have computed $K = 16$ different p-values. For simplicity, we assume that the data to compute each of them are different (independent).

$$\text{FWER} = \text{P}(\text{at least one of the } K \text{ p-values are significant})$$
$$= 1 - \text{P}(\text{none of the } K \text{ p-values are significant})$$
$$= 1 - \text{P}(\text{1st is not significant}) \times \cdots \times \text{P}(K\text{-th is not significant})$$
$$= 1 - (1 - 0.05) \times \cdots \times (1 - 0.05) \quad \text{(as no association exists)}$$
$$= 1 - (1 - 0.05)^K$$

K	1	2	3	4	5	10	**16**	20	50
FWER (%)	5	10	14	18	23	40	**56**	64	92

Cartoon: 56% chance of at least one significant false finding if no association exists.

FWER control

When we plan to compute $K \geq 1$ p-values, we can adjust their computation to control the FWER.

Bonferroni adjustment:

▶ adjusted p-value$= K \times$ original p-value

▶ adjusted significance level $= \alpha/K$.[12]

---

[12]Can be used to compute adjusted confidence intervals.

[13]Not allowed to keep testing until one significant result pops up and then divide all p-values by the number of tests performed.

FWER control

When we plan to compute $K \geq 1$ p-values, we can adjust their computation to control the FWER.

Bonferroni adjustment:

▶ adjusted p-value$= K \times$ original p-value

▶ adjusted significance level $= \alpha/K$.[12]

Intuition:

▶ equally share/split the original significance level α between the tests.

▶ the "total" risk of error (FWER) cannot exceed the sum of the errors of each test.

Remarks:

▶ always works: no specific assumption.

▶ but only works if we prespecify the analysis with K tests.[13]

---

[12]Can be used to compute adjusted confidence intervals.

[13]Not allowed to keep testing until one significant result pops up and then divide all p-values by the number of tests performed.

Prespecification matters
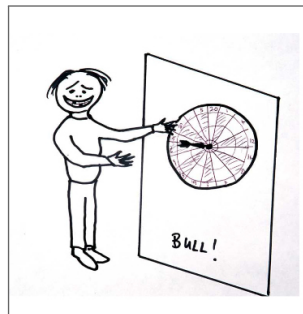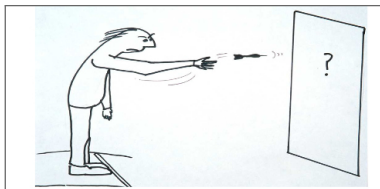
Concluding significance without prespecification is like drawing a dart-board around where the dart lands.

Bonferroni-Holm adjusted p-values

1. sort the p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$
2. adjust the first as with Bonferroni, i.e. $\tilde{p}_{(1)} = K \cdot p_{(1)}$ and others as

$$\tilde{p}_{(i)} = \min\left\{\tilde{p}_{(i-1)}, (K - i + 1) \cdot p_{(i)}\right\}$$

($\approx$ multiply the 1st by K, the 2nd by $K - 1$, the 3rd by $K - 2$, ...)

Remarks:

▶ same as for Bonferroni.

▶ we cannot compute corresponding adjusted significance levels and adjusted confidence intervals.

▶ less conservative than Bonferroni, i.e. adjusted p-values are always smaller.

Example

We compare 6 doses of treatments (10-60 mg) to placebo (0 mg).

Comparison	10 mg	20 mg	30mg	40mg	50mg	60mg
Original p-value	**0.005**	**0.009**	0.1	0.15	0.3	0.6
Bonferroni	**0.03**	0.054	0.6	0.9	1	1
Bonferroni-Holm	**0.03**	**0.045**	0.4	0.45	0.6	0.6

Note: we "truncate" the p-value to 1.

FWER vs FDR (1/2)

Controlling the FWER is important in "confirmatory" studies.

▶ When there is a clear prespecified scientific hypothesis and the aim is to "prove" it. E.g. clinical trial.

Controlling the FDR is often better suited in "exploratory" studies.

▶ When nice data are available, but no specific research questions / scientific hypotheses. You want to look at many associations and report findings which are "likely enough" true findings. E.g. Genomics.

False discovery rate (FDR): expected proportion of falsely rejected hypotheses among the rejected hypotheses.

Digression: confirmatory versus exploratory research

Exploratory research	Confirmatory research
No hypothesis required/hypothesis can be vague	Clear hypothesis required
Generate new hypothesis from data	Test *a priori* hypothesis with new data
High sensitivity desired, i.e. minimising the risk of false negatives	High specificity desired, i.e. minimising the risk of false positives
Suitable for making new discoveries and finding the unexpected	Suitable for establishing strong evidence and confirming the expected
For example: Testing of new compounds in mice	**For example:** Assessing the efficacy of a drug in humans

"It is essential to distinguish between exploratory and confirmatory research, and they are equally important to the scientific enterprise. Finding the questions to ask is at least as crucial as answering them, if not more so. But when the two concepts get confused, and the two worlds collide, the fallout can be disastrous"[14]

14 Schwab & Held. "Different worlds Confirmatory versus exploratory research." (2020) Significance, 8:9.

FWER vs FDR (2/2)

Hypotheses	Not rejected	Rejected	Total
True	U	V	K_0
False	T	S	$K - K_0$
Total	W	R	K

- $FWER = \mathsf{P}(V > 0)$
- $FDR = \mathsf{E}(V/R)$ (where here we set $V/R = 0$ if $R = 0$).
- controlling the FDR is less conservative than controlling the FWER: p-values adjusted to control the FDR are smaller than those adjusted to control the FWER.
- See Benjamini-Hochberg (1995) method to control FDR at e.g. 5%.

Outline/Intended Learning Outcome (ILOs)

Hypothesis testing
    ILO: to describe the principles and logic of hypothesis testing

One and two sample tests for continuous responses: t-test
    ILO: to identify when, how and why to use a t-test
    ILO: to define a p-value and contrast its use with that of a confidence interval

Power and Sample size calculation
    ILO: to identify why and how to make power and sample size calculations
    ILO: to analyse their strengths and limitations

Multiple testing
    ILO: to describe the multiple testing problem and employ basic remedies

Nonparametric test: Wilcoxon
ILO: to contrast pros and cons of Wilcoxon vs t-test

Wilcoxon-Mann-Whitney Test: motivation

Limitation of the two-sample t-test:

▶ Data should be normally distributed in each group

▶ **OR** the sample size of each group should be large.

Challenge:
What if we want a reliable computation of a p-value to compare two
groups, **with small sample data not necessarily normally distributed**?

A solution:
We can use a rank-based test[15]: the Wilcoxon-Mann-Whitney test[16]. It
provides "exact" p-values.[17]

Another advantage of Wilcoxon is its "robustness" to outliers, which
might be convenient.

_____

[15]also often called "non-parametric" test

[16]sometimes just called "Wilcoxon" or "Mann-Whitney" test.

[17]exact means that p-values are always valid (i.e. no "large n" approximation.)

Case: gene expression

▶ **Research question:**
Is the length of the candidate gene NACP associated with the level of expressed alpha synuclein mRNA, which has been shown to be associated with alcoholism?

▶ **Outcome:** level of expressed alpha synuclein mRNA.

▶ **Compared groups:** "short" vs "long" allele length (sum score built from additive dinucleotide repeat length categorized into groups).
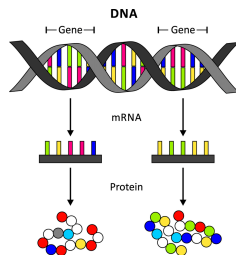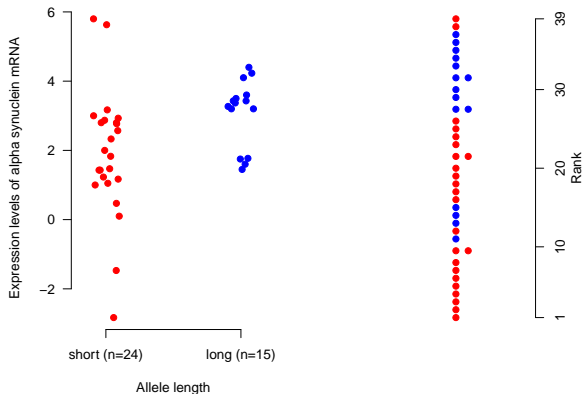

Challenges:

   ▶ small sample size $n = 24$ (short) $+$ 15 (long)

   ▶ outcome not known to be normally distributed.

   ▶ aim to **confirm** that this gene is linked to alcohol dependence.

Wilcoxon test: example

p-value=0.002

Why using the ranks:
If the two groups are similar, then the ranks should be equally distributed between the two groups. Whatever the distribution of the observations in each group, a randomly drawn blue observation should be larger than a randomly drawn red observation in about 50% of the draws. (Here $P(X > Y)$= 79.2 %)

Wilcoxon test: practical limitation

When a significant difference is shown **we can conclude that the distribution in the two groups are different, but nothing else**... which can be frustrating.

Common error/overinterpretation: conclude to a difference in median.

We cannot estimate a nice matching 95% CI to quantify the "effect size". By contrast, to complement the p-value of a t-test we can provide a matching 95% CI of the difference in mean.

Hence unless an "exact" p-value computation is really needed, using a t-test, possibly after having transformed the data, can often be preferred[18].

---

[18]See e.g. le Cessie, Goeman, and Dekkers. "Who is afraid of non-normal data? Choosing between parametric and non-parametric tests." European Journal of Endocrinology (2020).