Uses and abuses of tests

- Report the P-value
- Report a confidence interval
- Consider the model
- Consider the study design
- Be careful about data snooping

Was the result significant?

- In genetics, people often talk about
  - “suggestive” $5\% < P < 10\%$
  - “significant” $1\% < P < 5\%$
  - “highly significant” $P < 1\%$

  I despise this!

- Hard-and-fast rules are bad
  
  $P = 4.8\%$ is essentially the same as $P = 5.3\%$.

- Give the actual P-value, and treat it as a measure of evidence.
Was the result important?

- **Statistically significant** is not the same as **important**.
- A difference is “statistically significant” if it cannot reasonably be ascribed to chance variation.
- With lots of data, small (and unimportant) differences can be statistically significant.
- With very little data, quite important differences will fail to be significant.
- **Always report a confidence interval!**

Consider: \(0.5 \pm 0.1\) vs. \(100 \pm 40\)

---

Failure to reject

- **Failure to reject** the null hypothesis does not mean you should **accept** the null hypothesis.
- The means of two populations can always reasonably be **slightly** different—it’s impossible to prove, “They are the same,” though we can say, “They are not too different.”
- Think about the **power** of the statistical test.
- **Look at the confidence interval.**
Statisticians as cops

- Don’t think of statistics as a barrier to publishing important work.
- Rather, think of statistics as help for avoiding publishing garbage.
- Statistics can help you to avoid wasting time (and money) following false leads.

The role of the model

- Statistical tests and confidence intervals concern inferences about a (possibly hypothetical) population on the basis of data.
- **Model**: $X_1, \ldots, X_n$ independent with mean $\mu$ and SD $\sigma$.
- For a well-designed (randomized) experiment, this is usually not a worry.
- Be suspicious about statistical tests with censuses and convenience samples.
Does the difference prove the point?

- A test of significance doesn’t check the design of the study.
- With observational studies or poorly controlled experiments, the proof of statistical significance may not prove what you want.
- **Example:** consider the tick/deer leg experiment. It may be that ticks are not attracted to deer-gland-substance but rather despise the scent of latex gloves and deer-gland-substance masks it.
- **Example:** In a study of gene expression, if cancer tissue samples were always processed first, while normal tissue samples were kept on ice, the observed differences might not have to do with normal/cancer as with iced/not iced.
- Don’t forget the science in the cloud of data and statistics.

**Data snooping / Multiple testing**

- Generally we perform more than one statistical test at once.
- If you are performing many statistical tests, and then reporting the interesting ones, take care!
  You need to adjust for the fact that you are performing many tests.
- Sometimes investigators study their data, and then apply formal statistical tests only to features that appear interesting (and likely statistically significant).
  **Take care!** They should adjust for the statistical tests that they applied informally, in snooping through their data.
- Ideally, such multiple statistical tests are treated as exploratory, and the interesting results are confirmed with independent data.
Backcross experiment

Phenotype distributions

- Within each of the parental and F₁ strains, individuals are genetically identical.

- Environmental variation may or may not be constant with genotype.

- The backcross generation exhibits genetic as well as environmental variation.
Data and Goals

**Phenotypes:** \( y_i \) = phenotype for mouse \( i \)

**Genotypes:** \( x_{ij} = 1/0 \) if mouse \( i \) is BB/AB at marker \( j \)
(for a backcross)

**Genetic map:** Locations of markers

**Goals:**
- Identify the (or at least one) genomic regions (QTLs) that contribute to variation in the phenotype.
- Form confidence intervals for QTL locations.
- Estimate QTL effects.
The simplest method: t-tests

- Split mice into groups according to genotype at a marker.
- Do a t-test
- Repeat for each marker.
Adjustment for multiple tests

- We performed a t-test at each of 91 markers. (The markers are, of course, associated.)

- The maximum t-statistic was 3.05. What P-value do we assign to this?

  Nominal P-value = Percentile of $|T|$ (under null hypothesis) = 0.002

  **Adjusted P-value = Percentile of maximum $|T|$ (under null hypothesis of no QTLs anywhere)**

- How to get at the distribution of the maximum $|T|$, genome-wide? I like permutation tests. They require heavy computation, but they're trustworthy.
Permutation tests

- Permute/shuffle the phenotypes; keep the genotype data intact.
- Calculate $|T^*(z)| \rightarrow M^* = \max_z |T^*(z)|$
- We wish to compare the observed $M$ to the distribution of $M^*$.
- $\Pr(M^* \geq M)$ is a genome-wide P-value.
- The 95th %ile of $M^*$ is a genome-wide critical value
- We can’t look at all $n!$ possible permutations, but a random set of 1000 is feasible and provides reasonable estimates of P-values and critical values

Permutation distribution

Observed max $|T|$

Area to right = 7%
Uses and abuses of tests

- Report the P-value
- Report a confidence interval
- Consider the model
- Consider the study design
- Be careful about data snooping