Lecture 9.1

Hypothesis testing (continued...)

Power, Type I, and Type II errors

It is possible that we will reject H_0 when H_0 is in fact true. This is called a **Type I error** whose probability is controlled by the significance level.

The other error we can make in hypothesis testing is to 'accept' H_0 when it is false. This is called a **Type II error**.

The **power** of the test is given by $1 - \mathbb{P}(\text{Type II Error})$. Power is affected by many factors including the size of the effect under H_1, population variance, sample size and significance level.

Power, Type I, and Type II errors

	Decision	
	Retain	Reject
H_0 is true	Correct $(1-\alpha)$	Type I Error (α)
H_0 is false	Type II Error (β)	Correct Power $= (1-\beta)$

Ideally, we would like the probability of making Type I and Type II errors to be as small as possible. However, there is a trade-off between the two errors. Intuitively, when the significance level, α, is decreased (smaller probability of type I error), it is more likely that p-value $> \alpha$ (so accept H_0 even when it is false).

Example

An automobile model is known to sustain no visible damage 25% of the time in 10-mph crash tests. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the proportion of all 10-mph crashes with this new bumper that result in no visible damage. The hypotheses to be tested are

$$H_0 : p = 0.25, \quad \text{against} \quad H_1 : p > 0.25.$$

Example (continued)

The test will be based on an experiment involving $n = 20$ independent crashes with prototypes of the new design.

Let X be the number of crashes with no visible damage (test statistic).

Under the null hypothesis, $X \sim \text{Bin}(20, 0.25)$, the p-value for a given observed value x is

$$\mathbb{P}(X \geq x) = \sum_{i=x}^{20} \binom{20}{i} \times 0.25^{i} \times 0.75^{20-i}.$$

So, we have

$$\mathbb{P}(X \geq 7) = 0.214, \quad \mathbb{P}(X \geq 8) = 0.102, \quad \text{and} \quad \mathbb{P}(X \geq 9) = 0.041.$$

Example (continued)

Consider using a significance level of 0.05. Thus, rejecting H_0 when p-value ≤ 0.05 is equivalent to rejecting H_0 when $X \geq 9$.

$$P(\text{committing a type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$$
$$= \mathbb{P}(X \geq 9 \text{ when } X \sim \text{Bin}(20, 0.25)) = 0.041 \leq 0.05.$$

That is, the probability of a type I error is controlled by the significance level.

If the null hypothesis is true and the test procedure is used over and over again, each time with a group of 20 crashes, in the long run the null hypothesis will be incorrectly rejected in favor of the alternative hypothesis about 4% of the time.

Example (continued)

There is only one type I error probability because there is only one value of the parameter for which H_0 is true, i.e., $p = 0.25$.

Let β denote the probability of committing a type II error. Unfortunately there is not a single value of β, because there are a multitude of ways for H_0 to be false: it could be false because $p = 0.30$, $p = 0.37$, $p = 0.05$, and so on. There is in fact a different value of β for each different value of p that exceeds 0.25.

At the chosen significance level 0.05, H_0 will be rejected if and only if $X \geq 9$, so H_0 will not be rejected if and only if $X \leq 8$.

Example (continued)

What is the probability of committing type II error when $p = 0.3$?

$$\beta(0.3) = \mathbb{P}(\text{type II error when } p = 0.3) = \mathbb{P}(H_0 \text{ is not rejected when } p = 0.3)$$
$$= \mathbb{P}(X \leq 8 \text{ when } X \sim \text{Bin}(20, 0.3)) = 0.887.$$

When p is actually 0.3 rather than 0.25 (a "small" departure from H_0), roughly 89% of all experiments of this type would result in H_0 incorrectly standing! This is because the sample size of 20 is too small to permit accurate discrimination between .25 and 0.3. The departure from H_0 needs to be larger for it to be detected with such a small sample size.

p	0.3	0.4	0.5	0.6	0.7	0.8
$\beta(p)$	.887	.560	.251	.056	.005	.000

Intuitively, the greater the departure from H_0, the more likely it is that such a departure will be detected.

Example (continued)

In order to detect small departures from H_0, we need larger samples. Let's $n = 1000$ in our example. Then, under the null hypothesis, $X \sim \text{Bin}(1000, 0.25)$, and the p-value for a given observed value x is

$$\mathbb{P}(X \geq x) = \sum_{i=x}^{1000} \binom{20}{i} \times 0.25^i \times 0.75^{1000-i}.$$

So, we have

$$\mathbb{P}(X \geq 272) = 0.059, \quad \mathbb{P}(X \geq 272) = 0.051, \quad \text{and} \quad \mathbb{P}(X \geq 274) = 0.044.$$

So, now what is the probability of committing type II error when $p = 0.3$?

$$\begin{aligned}\beta(0.3) &= \mathbb{P}(\text{type II error when } p = 0.3) = \mathbb{P}(H_0 \text{ is not rejected when } p = 0.3) \\ &= \mathbb{P}(X \leq 273 \text{ when } X \sim \text{Bin}(1000, 0.3)) = 0.033.\end{aligned}$$

How to chose H_0 vs H_1?

There is an asymmetry between the null and alternative hypotheses. The decision as to which is the null and which is the alternative hypothesis is not a mathematical one, and depends on scientist context, custom, and convenience. **Some guidelines are:**

- When one of the competing hypotheses is more complex, null hypothesis is chosen as the one which is simpler than the alternative,
- The consequences of incorrectly rejecting one hypothesis may be graver than those of incorrectly rejecting the other. In such a case, the former should be chosen as the null hypothesis, because the probability of falsely rejecting it, i.e., Type I error, could be controlled. For example, in scientific studies, the false confirmation of one's own theory (Type I error) is typically a more serious error than falsely failing to confirm ones' own theory (Type II error).

Caution!

- The p-value resulting from carrying out a test on a selected sample is not the probability that H_0 is true.
- Care must be taken in interpreting evidence when the sample size is large, since any small departure from H_0 will almost surely be detected by a test as **statistically significant**, yet such a departure may not really be **practically significance**.
- H_0 not only includes the assumptions about the parameters, but it also contains assumptions about the underlying distribution of the data. Small p-value implies inconsistency with *all* of our assumptions, i.e., perhaps our initial assumption about the model for the distribution of the data was altogether wrong!

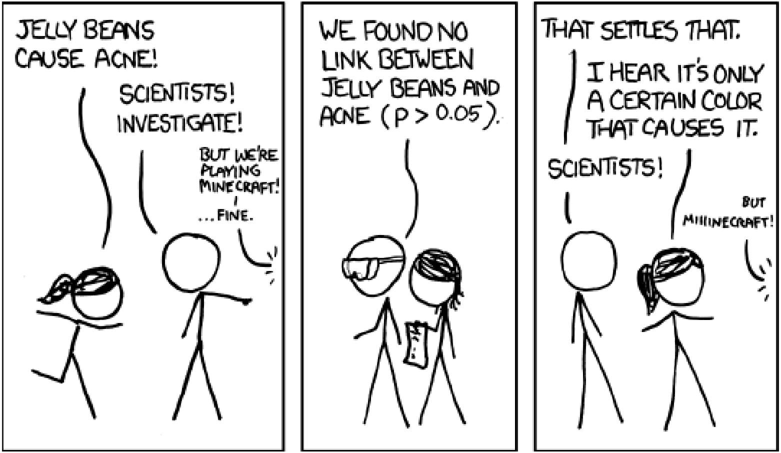
Paired data

Analysis of Paired Data

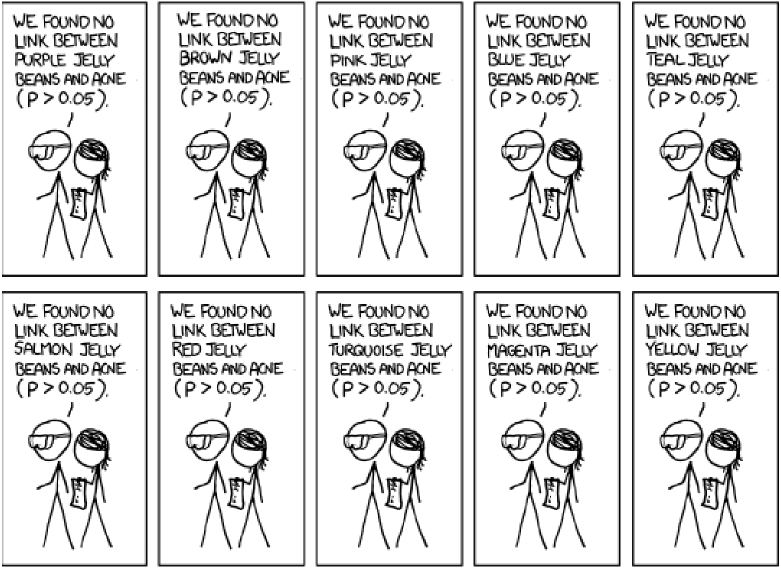
Our two-sample testing techniques require the assumption that all X and Y samples are independent of one another. In many experiments, there is only one set of n experimental objects; making two observations on each one results in a natural pairing of values. Such data often arises in "before–after" experiments, e.g., X_i and Y_i represent, respectively, the "before" and the "after" status of the same the i^{th} object.

To compare the difference in the expectations of two dependent random variables X and Y, based on paired samples $\{X_i\}$ and $\{Yi\}$, we use the difference random variable $D = X - Y$, and test whether $\delta = \mu_X - \mu_Y = 0$ using independent samples $D_i = X_i - Y_i$. We are thus back to the case of a one-sample testing.

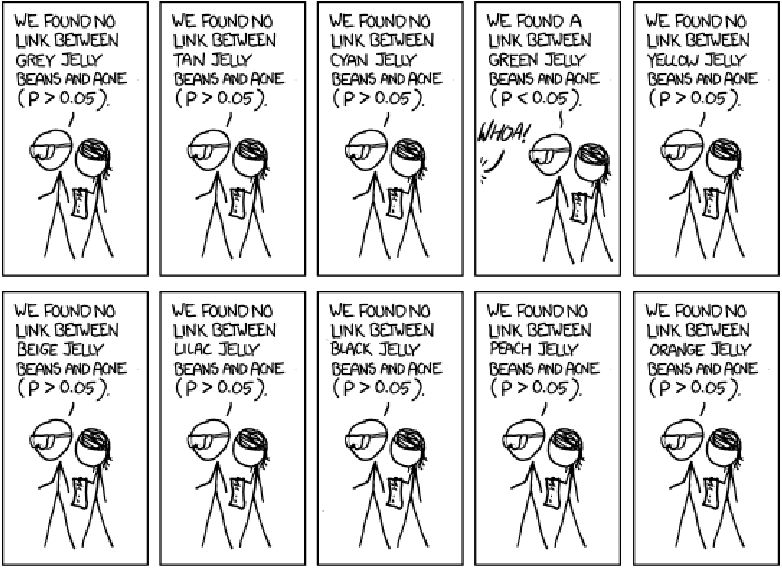
Multiple hypothesis tests



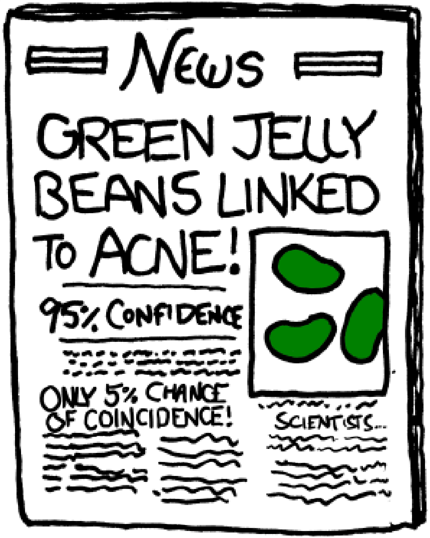
Multiple hypothesis tests



Multiple hypothesis tests



Multiple hypothesis tests



Multiple hypothesis tests

Simultaneous testing of Several Hypotheses:

When doing k hypothesis testing, we need to set the significance level of each test to α/k so that overall, the probability of making at least one type I error, i.e., overall significance level, remains below α.

This is because by *Bonferroni inequality*, we have

$$\begin{aligned}
\mathbb{P}(\text{at least one type I error}) &= \mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_k) \\
&\leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_k) \\
&= k\alpha/k = \alpha.
\end{aligned}$$

Moral of the story: Does our assumptions match our observations?

- Yes: Keep our assumptions
- No: Discard our assumptions