Types of errors in a hypothesis test

		Truth	
		H_o **true**	H_A **true**
Decision	**Accept** H_o	*correct decision, 1-*α	Type II error $= \beta$
	Reject H_o	Type I error $= \alpha$	*correct decision, 1-*β

Evaluating a test procedure

Evaluating of hypothesis testing procedure involves consideration of:

- What is the effect size we are trying to detect under the hypotheses?
- What errors (Type I, Type II) are we willing to tolerate?
- How much variability is present? Will we be able to detect the effect?
- What sample size is required?

Power analysis: *Before collecting data,* evaluate the power of a hypothesis test to detect a significant result when one exists.

Example

The EPA sets a limit of 5ppm on PCB in water. A major manufacturing firm producing PCB for electrical insulation discharges small amounts from the plant. The company management, attempting to control PCB discharge levels, has given instructions to halt production if the mean amount of PCB in the effluent exceeds 3ppm.

Their test: $H_o : \mu \leq 3$ ppm; $H_A : \mu > 3$ ppm

Suppose that data have been collected and that the null hypothesis is *not* rejected. Results:

- No action is taken to reduce discharge of the pollutant.
- A report is printed that says, *"Our statistical analysis has shown that our company is in compliance."*

Perspectives on possible errors: EPA, manager of the company producing the pollutant, nearby residents, conservation biologist

Power Defined

- The chance of reporting a statistically significant difference when the treatment really produces an effect
- Power=1-Type II error=$1 - \beta$
- Defined with respect to a *specific* alternative hypothesis being considered.
- If an hypothesis test has power equal to 0.8, there is an 80% chance of actually reporting a statistically significant effect when one is present.

What factors affect power?

The ability to detect an effect (such as a non-zero difference in means) with a given level of confidence depends on:

1. Size of the difference or treatment effect
2. Sample size
3. Inherent variability in the data
4. The level (α) of the Type I error you are willing to tolerate. (It is easier to find a difference if you take a bigger chance on a false positive.)

PCB in Effluent Example

Elements of Hypothesis Testing Procedure Chosen by the Company:

1. Claim is that the company is out of compliance. $H_o : \mu \leq 3$ ppm; $H_A : \mu > 3$ ppm
2. Set α=0.01.
3. Set $n = 30$
4. From previous analyses, $s = 0.5$ ppm. Taken to be known.

What if in fact the true mean is 3.2 ppm? Would our test as defined above have sufficient power to detect this?

Calculating Power of a Test: Steps

1. Specify the hypothesis test under consideration.
2. Set the Type I error.
3. Define σ and assume known.
4. Define rejection region for the test.
5. Define the *non-rejection region* for the test.
6. Rewrite this region *in terms of the sample mean (or difference of sample means)*
7. Now specify the alternative hypothesis: $H_A : \mu = 3.2$.
8. Calculate β, the Type II error, the probability of accepting the null if the alternative is true.
9. Power $= 1 - \beta$

Power: Another Example

An experiment is planned to assess the effects of automobile pollution on plants. Twenty soil specimens will be taken 0.6 m from the roadside, and twenty soil specimens will be taken 6.0 m from the roadside, and their nitrogen concentrations will be measured.

Suppose that a previous study has established that $s_{pooled} = 0.8$, and thus we treat σ as known to be 0.8.

Researchers wish to determine whether there is sufficient evidence to conclude that the mean nitrogen concentration in soil 0.6 m from the roadside is higher than at 6 m. They have set $\alpha = 0.01$.

The researchers wish to detect an increase of up to 0.4 mg/g in nitrogen levels nearby (0.6 m) over those farther away (6.0 m).

What would be the Type II error and the power for this test? What would you suggest to increase the power in this case?

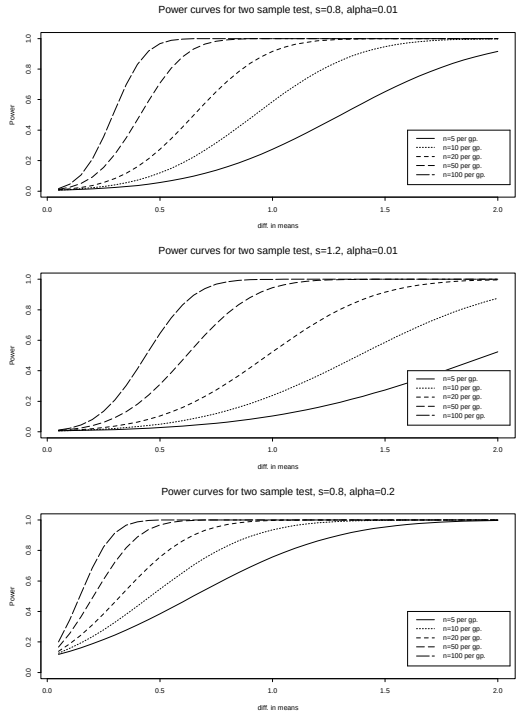
Required sample size and power

A Large sample approximation to the minimum sample size to achieve power of $1 - \beta$:

$$n \geq a \left(z_{1-\beta} + z_{1-b\alpha}\right)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- δ is the effect size (magnitude of deviation from null).
- $a = 1$ for one-sample test.
- For a two-sample test, $a = 2$. Formula assumes equal sample sizes of n in each group.
- $b = 1$ for one sided test; $b = 1/2$ for two-sided test.
- Explain intuitively why the sample size for $\delta = 8$ and $\sigma = 2$ is the same as the sample size for $\delta = 4$ and $\sigma = 1$ under this formula.
- Ref: Rao, *Statistical Methods for the Life Sciences*

Power Curves



Prospective Power Analyses

- determine number of replicates or samples needed to achieve power (1-β) for given δ, α, σ.
- cost or logistical constraints on number of samples given δ, α, σ, β.
- minimum effect size (δ) to detect given target α, σ, n.
- Scheiner & Gurevich, in *Design and analysis of ecological experiments*, suggest sensitivity analysis for σ. May need to reduce σ through blocking or measuring other covariates.

Retrospective Power Analysis

- Situation: H_0 not rejected because...
  - true effect not biologically important, and null is "nearly true"
  - true effect size was biologically important but we failed to reject null.
- "Observed power" gives the same information as the p-value. Also common in literature: an after-the-fact "detectable effect size"
- Prefer more thought given to choice of hypotheses.
- Prefer CI's which give an idea of the bounds on the effect size. Are there biologically important effect sizes in these bounds? Which parameter values are supported by the data?