

Power and Size

QUANTIFYING THE ERRORS OF A HYPOTHESIS TEST

Two Experiments

Suppose we are running two chemistry experiments.

We run two-sample, two-sided, pooled t-tests for differences in the mean

Experiment 1:

We calculate a very low p-value (say 0.0001) and choose to reject the null hypothesis

Experiment 2:

We calculate a large p-value (say 0.2) and choose to fail to reject the null hypothesis

Of course, we don't know the truth. But, is there anything that can be said further about the types of errors we might be making?

Types of Errors

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

A Thought Experiment

Suppose we specify a null hypothesis of: $H_0: \mu_1 - \mu_0 = D = 0 \text{ mmHg}$

A clinically meaningful difference is: $\mu_1 - \mu_0 = D = 4 \text{ mmHg}$

(This is the smallest difference we want to be able to detect)

Can we say how likely our test is to detect this difference?

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

Power

The **POWER** of a hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is false ($1 - \text{Probability}(\text{Type II error})$)

Power generally increases with sample size

The test need be sufficiently powered to detect a meaningful difference (e.g. 4 mmHg in the example)

An increase in power means an increase in the cost of conducting the experiment (e.g. measuring more people's middle finger)

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

Power

The **POWER** of a hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is false

To compute the power of a test, we need:

1. A clinically meaningful effect size
2. The standard deviation (σ)
3. The total number of observations (n)

Important: Power calculations should be done **before** the data is gathered!

The Standard Deviation

The standard deviation (σ) must be specified to compute the power

Some ways of accomplishing this:

- Looking at other studies on the same/similar subject
- Ask a subject-matter expert for a range of reasonable values
- Run a small, preliminary experiment to estimate σ

(Technically, this data should not be used in later analysis, but practically speaking this preliminary data is often included)

Total # of Observations

The usual trade-off

- Power tends to increase as the number of observations (n) increase
- The cost and complexity of the study tends to increase with (n) as well

→ Choose n large enough, but no larger, to have an acceptable probability of detecting a clinically meaningful difference

SAS has a nice procedure for doing just that..

PROC POWER

Important: There is no data here!

```
PROC POWER;  
  TWOSAMPLEMEANS  
  MEANDIFF = 3.5 4 4.5  
  STDDEV= = 8 8.5 9  
  POWER = 0.8  
  NTOTAL = .;  
  PLOT Y=POWER min=0.5 max=0.99;  
RUN;
```

Look at a range of n

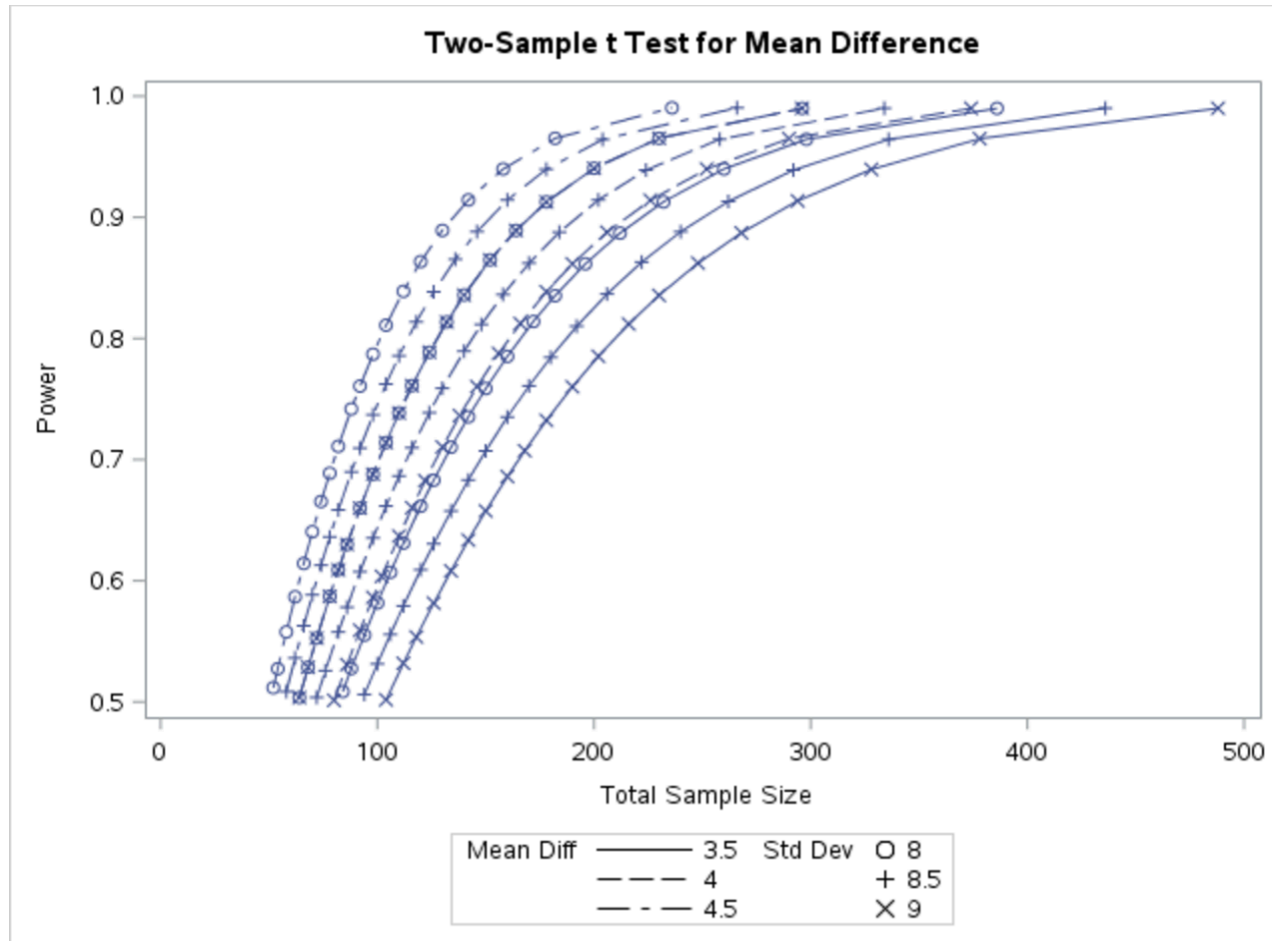
Plot power from .5 to .99 over a range of n (next slide)

The POWER Procedure Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total				
Index	Mean Diff	Std Dev	Actual Power	N Total
1	3.5	8.0	0.800	166
2	3.5	8.5	0.802	188
3	3.5	9.0	0.801	210
4	4.0	8.0	0.801	128
5	4.0	8.5	0.801	144
6	4.0	9.0	0.803	162
7	4.5	8.0	0.803	102
8	4.5	8.5	0.800	114
9	4.5	9.0	0.801	128

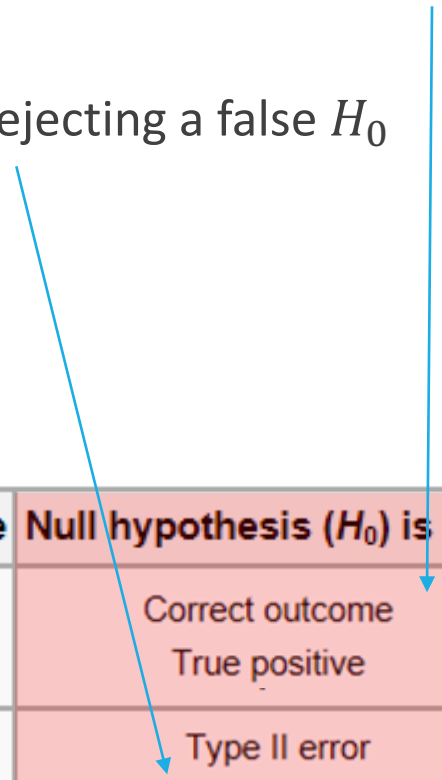
Power Plot



Complement of Power

The **POWER** of a hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is false

Hence, a Type II error is the probability of not rejecting a false H_0



	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

PROC POWER

Important: There is no data here!

```
PROC POWER;  
  TWOSAMPLEMEANS  
  MEANDIFF = 3.5 4 4.5  
  STDDEV= = 8 8.5 9  
  POWER = 0.8  
  NTOTAL = .;  
  PLOT Y=POWER min=0.5 max=0.99;  
RUN;
```

For these possible experiments, the
Type II error = 0.2

The POWER Procedure Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total				
Index	Mean Diff	Std Dev	Actual Power	N Total
1	3.5	8.0	0.800	166
2	3.5	8.5	0.802	188
3	3.5	9.0	0.801	210
4	4.0	8.0	0.801	128
5	4.0	8.5	0.801	144
6	4.0	9.0	0.803	162
7	4.5	8.0	0.803	102
8	4.5	8.5	0.800	114
9	4.5	9.0	0.801	128

Type I Error

- Now, consider what happens if H_0 is true
- **Experiment:** We run two-sample, one-sided, pooled t-test for differences in the mean. There are $n = 50$ subjects, so $df = 48$.
- Suppose we will reject H_0 if the p-value is below 0.01
- Hence, we will reject if
(observed t-stat) $> t_{\alpha=0.01,48} = 2.40658$ **(CRITICAL VALUE)**
- The value α is a Type I error and is known as the **SIZE** of the hypothesis test

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

Putting it All Together

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

