Section 6.4 - Power and Inference as a Decision

Statistics 104

Autumn 2004

Errors in Hypothesis Testing

	Reject H_0	Don't Reject H_0
H_0 true	✗(Type I Error)	✓
H_0 false	✓	✗(Type II Error)

Two types of errors

- Type I error: rejecting H_0 when you shouldn't (False positive)
- Type II error: don't reject H_0 when you should (False negative)

There is also (slightly humourous) proposed third error

- Type III error: solving the wrong problem precisely (attributed to Howard Raiffa) (Also referred to as a Type 0 error)

  It is better to come up with an approximate answer to the correct question than an exact answer to the wrong one.

Want to have hypothesis tests where the chance of either a Type I or II error is small.

$$P[\text{Type I error}] = \alpha \qquad \text{(Significance level)}$$

This is set by the researcher.

The chance of making a type II error depends on where in the alternative space we actually are.

Suppose that we want to test $H_0 : \mu = 0$. It seems reasonable that we would make the wrong conclusion when $\mu = 1$ more often than when $\mu = 1000$, all other things being equal.

In fact this can be justified.

The approach to justify statements like this is with the concept of power.

Power:

The probability that a fixed level α hypothesis test will reject H_0, when a particular alternative value of the parameter is true.

This relates to the probability of making a correct rejection.

Different parameter values give different power.

Note that Power $= 1 - P[\text{Type II error given } \mu = \mu_1]$

Calculating power:

1. State H_0, H_A, and α.

2. Determine which values of the test statistic will lead us to reject H_0.

3. Calculate the probability of rejecting H_0 when a particular alternative (μ_1) is true.

Lets examine the idea with the z-test.

Example: Comparing shoe sole material

2 materials A and B.

Want to know which material is better with respect to wear.

10 boys, each with special shoes. One shoe is made with material A and the other with material B.

Response is $X =$ A wear $-$ B wear.

1. Examine $H_0 : \mu = 0$ versus $H_A : \mu > 0$ with an $\alpha = 0.05$ test.

   Assume that σ for the wear difference is 2.5.

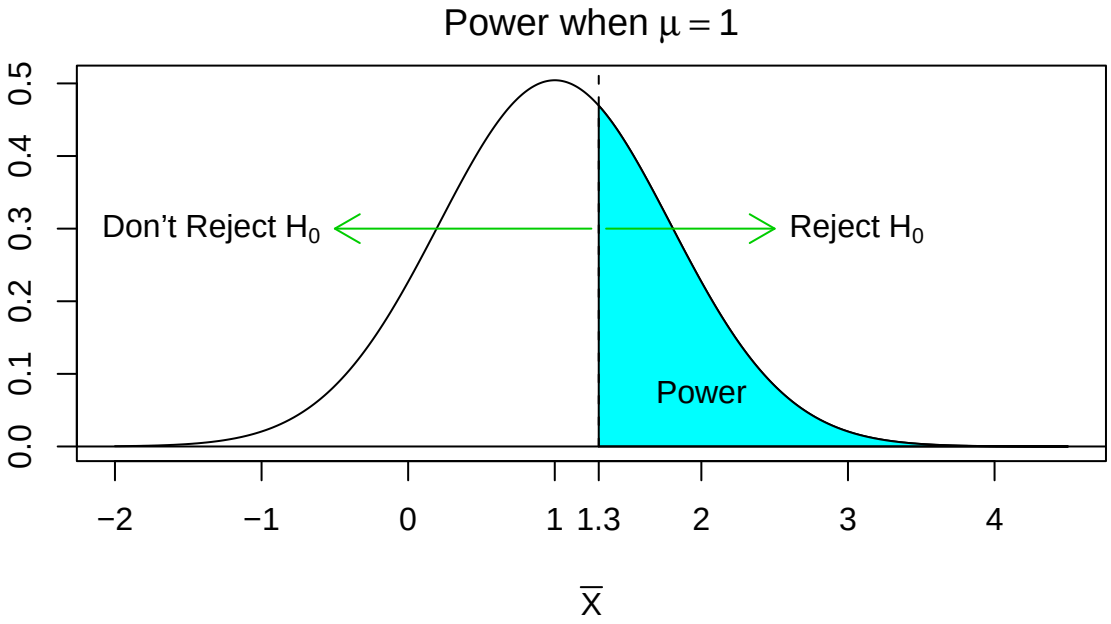
2. Will reject H_0 if

$$z = \frac{\bar{x}}{\frac{2.5}{\sqrt{10}}} = \frac{\bar{x}}{0.791} \geq 1.645$$

   or equivalently, if $\bar{x} \geq 1.300$.

3. Want power if real $\mu = 1$ $(= \mu_1)$. In this case $\bar{X} \sim N(1, 0.791)$ so we want

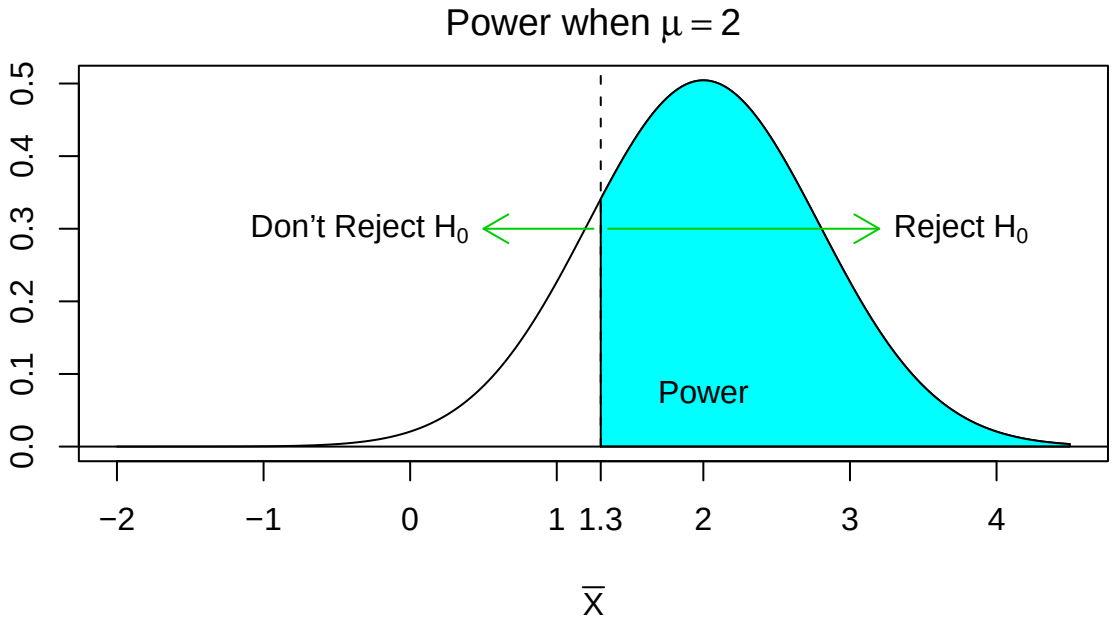
$$P[\bar{X} \geq 1.3 \text{ when } \mu = 1]$$



$$\text{Power}(\mu = 1) = P\left[\frac{\bar{X} - 1}{0.791} \geq \frac{1.3 - 1}{0.791}\right] = P[Z \geq 0.379] = 0.352$$

So we have a about a 35% chance of rejecting the null in this situation.

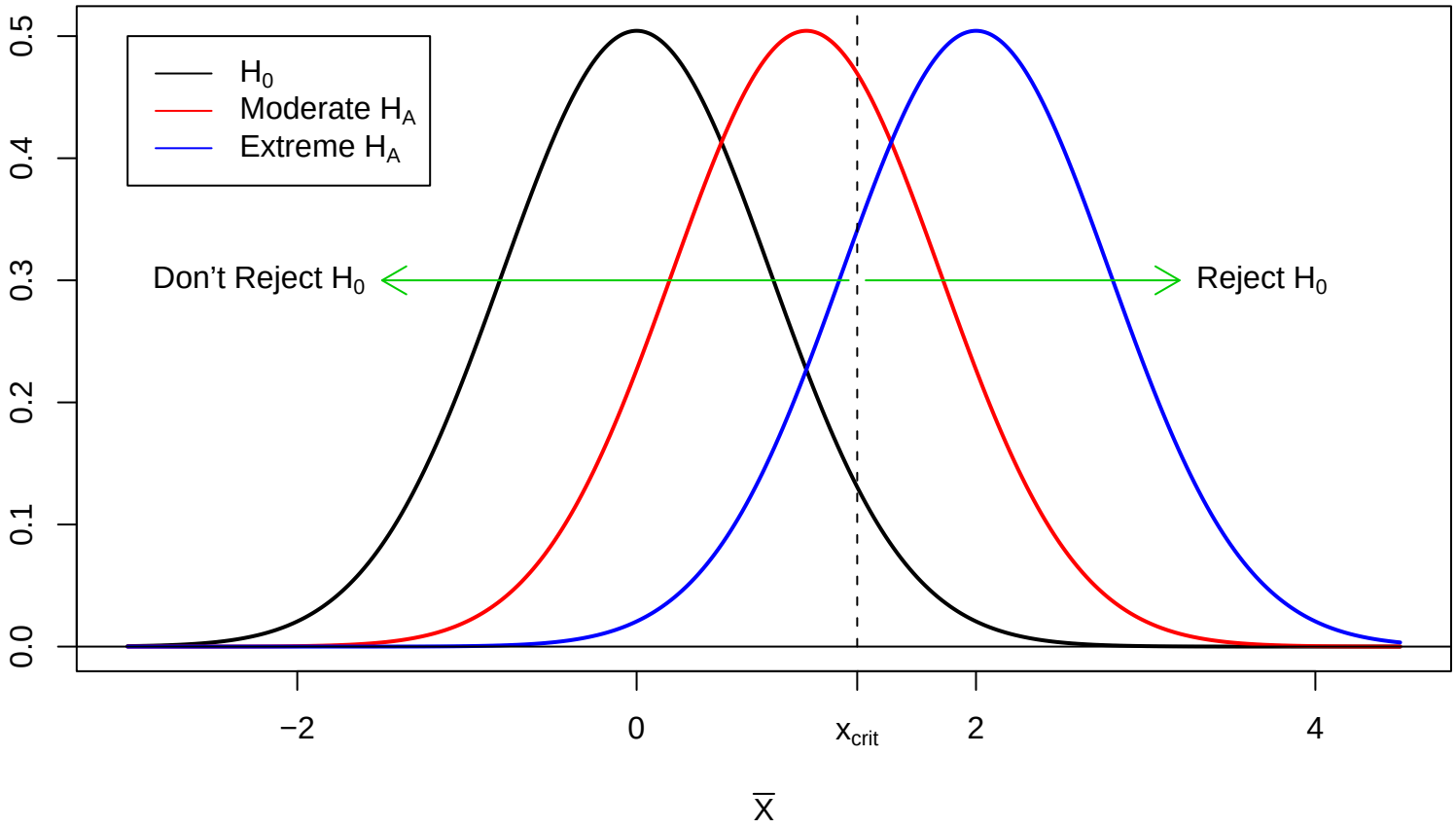
What about $\mu_1 = 2$. Want the $P[\bar{X} \geq 1.3 \text{ when } \mu = 2]$



$$\begin{aligned} \text{Power}(\mu = 2) &= P\left[\frac{\bar{X} - 2}{0.791} \geq \frac{1.3 - 2}{0.791}\right] \\ &= P[Z \geq -0.885] = 0.812 \end{aligned}$$

So we have a about a 81% chance of rejecting the null in this situation.

In general, as μ moves further into the alternative, the power increases.



What happens if the sample size changes while keeping μ the same. Lets look at the power when $n = 20$ and $\mu = 1$.

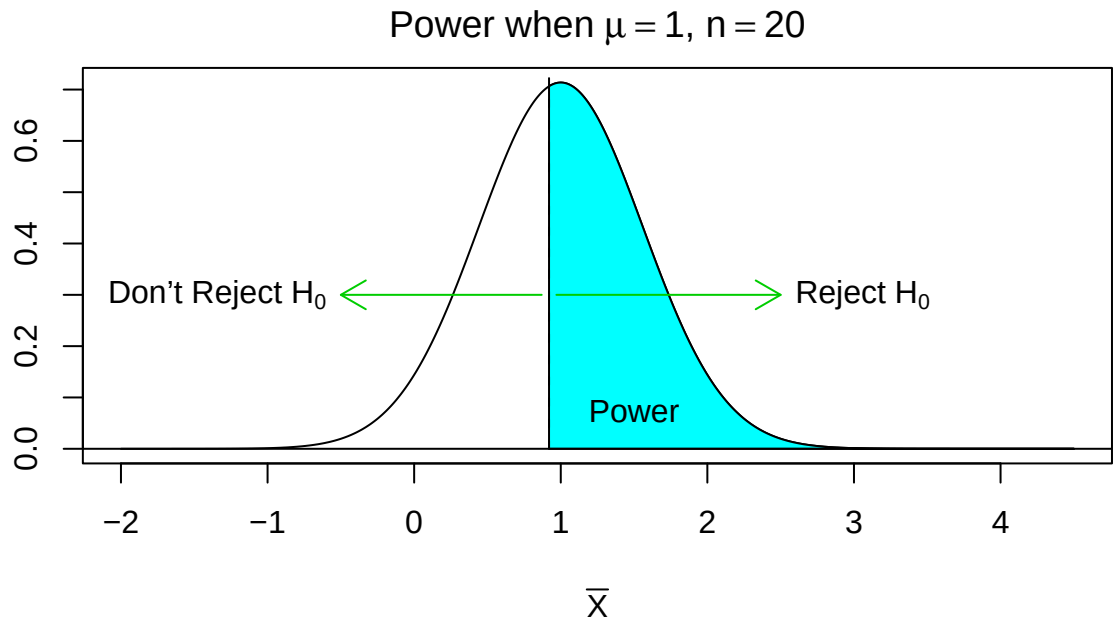
1. Examine $H_0 : \mu = 0$ versus $H_A : \mu > 0$ with an $\alpha = 0.05$ test.

2. Will reject H_0 if

$$z = \frac{\bar{x}}{\frac{2.5}{\sqrt{20}}} = \frac{\bar{x}}{0.559} \geq 1.645$$

   or equivalently, if $\bar{x} \geq 0.920$.

3. Want power if real $\mu = 1$ $(= \mu_1)$. In this case $\bar{X} \sim N(1, 0.559)$



$$\begin{aligned} \text{Power}(\mu = 1) &= P[\bar{X} \geq 0.92] \\ &= P\left[\frac{\bar{X} - 1}{0.559} \geq \frac{0.92 - 1}{0.559}\right] \\ &= P[Z \geq -0.143] = 0.557 \end{aligned}$$

So increasing the sample size from 10 to 20 in this case increases the power from 0.352 to 0.557.

In general, increasing the sample size will increase power. This can be seen by looking at the critical region of the test. For a 5% one sided test of $H_0 : \mu = 0$, the rejection region is

$$\bar{x} \geq \frac{1.645\sigma}{\sqrt{n}}$$

This decreases as n increases, implying for bigger sample sizes there are more possible sample averages that will be declared statistically significant. This come from the concentration of the sampling distribution about its mean as n increases.

For a two-sided z-test, the approach is similar, but you need to deal with both tails. For example, for a 5% two-sided test for the earlier situation ($n = 10$), the rejection region is to reject H_0 if

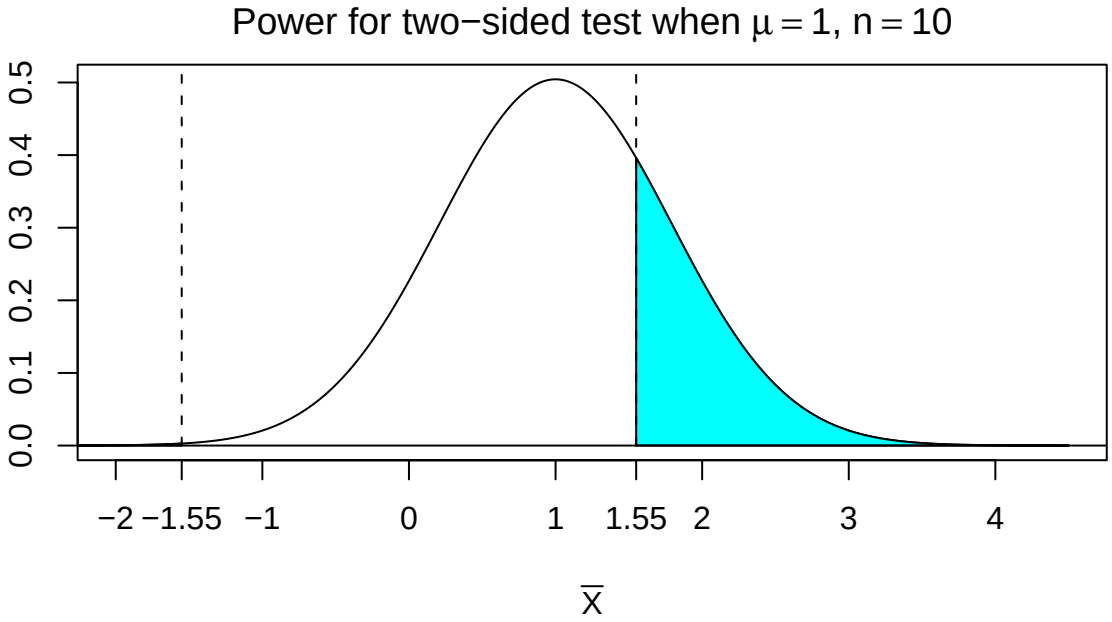
$$z \leq -1.96 \quad \text{or} \quad z \geq 1.96$$

For $n = 10$, this corresponds to

$$\bar{x} \leq -1.550 \quad \text{or} \quad \bar{x} \geq 1.550$$

So the power in this case is

$$P[\bar{X} \leq -1.55 \text{ when } \mu = 1] + P[\bar{X} \geq 1.55 \text{ when } \mu = 1]$$



Power is often used as an approach to determining sample sizes in studies. The idea is to figure out what sample size will give you a certain power (say 80%), for an hypothesis of interest (say $\mu = \mu_1$) with a test with a given significance level α.

So for example, find n satisfying

$$P\left[\bar{X} \geq \frac{1.645\sigma}{\sqrt{n}} \text{ when } \mu = 1\right] = 0.8$$

In this case n needs to be at least 39.

When designing studies, you need to choose reasonable values for power, α, and the alternative value μ_1. Choosing a sample size that gives a desired power when $\mu = 10$ when in fact your are looking for effects around $\mu = 1$ is likely to lead to a too small a sample size and an underpowered study.

How to Increase Power:

1. Increase n. More data will provide more information to distinguish among different parameter values.

2. Change your sampling/experimental procedure. Other designs may lead to smaller standard errors, which acts like increasing your sample size or decreasing σ.

3. Consider different alternatives. More extreme alternatives have higher power. Note this is more a consideration when it comes to choosing a sample size. While you might not be able to reach a desired power with say $\mu = 1$, you might be able to do it with $\mu = 2$.

4. Increase α. With a larger α, you have a larger critical region, so it is easier to reject.

For example, for a two-sided z-test,

α	Rejection Region
0.05	$\lvert z\rvert \geq 1.960$
0.10	$\lvert z\rvert \geq 1.645$

5. Decrease σ. This has a similar effect to increasing the sample size. Improving the measurement procedure and focusing on subpopulations are two approaches to this.

These are the same sort of ideas proposed when trying to get a narrower confidence interval.