STA261: Week 10

Power and Sample Size Calculations

Alex Stringer

March 12th - 16th, 2018

Disclaimer

The materials in these slides are intended to be a companion to the course textbook, *Mathematical Statistics and Data Analysis, Third Edition*, by John A Rice. Material in the slides may or may not be taken directly from this source. These slides were organized and typeset by Alex Stringer.

A big thanks to Jerry Brunner as well for providing inspiration for assignment questions.

License

Recap

So far we have talked about a variety of types of Hypothesis Tests:

- Normal theory (justified by CLT)
- Likelihood Ratio
- Two sample: independent and paired

We said for all of these: let's fix the test to have a controlled probability of Type I error.

Power

Recall: Type I error $=$ Reject H_0 when it is true.

We haven't talked at all about Type II error: fail to reject H_0 when it is false.

What we are interested in is actually $1 - P(\text{Type II error})$, the probability of rejecting a false null hypothesis.

This is in some sense a measure of how sensitive the test is.

Power

Definition: the **power** of a hypothesis test, η, is

$$\eta = P(\text{Reject}H_0 | H_0 \text{ false}) = 1 - P(\text{Type II error})$$

We want our test to have high power. Tests with higher power are able to detect smaller deviations from H_0, and are therefore "stronger".

Composite Alternatives

When we wanted to calculate the significance level of our test, it was straightforward, because under H_0, the distribution of the test statistic $T(\mathbf{X})$ was specified completely- we considered only **simple** null hypotheses.

Calculating power is more difficult, because we consider **composite** alternatives.

The distribution of the test statistic when H_0 is false depends on *how false it is*.

Z-Test Power

Let's calculate the power of the standard normal-theory Z-test.

$X_i \sim N(\mu, \sigma_0^2)$: IID random sample from a normal distribution with unknown mean and variance.

We said we reject $H_0 : \mu = \mu_0$ if

$$\left| \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \right| > z_{1-\alpha/2}$$

which if H_0 is true, happens with probability α.

Z-Test Power

What if H_0 is false, and $\mu = \mu_1 \neq \mu_0$?

The power is the probability of rejecting H_0 in this situation.

$$\begin{aligned}\eta &= P\left(\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right| > z_{1-\alpha/2}\right) \\ &= 1 - P\left(-z_{1-\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} < z_{1-\alpha/2}\right)\end{aligned}$$

Z-Test Power

$$
\begin{aligned}
\eta &= 1 - P\left(-\frac{\mu_1}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} < \frac{\bar{X} - \mu_0 - \mu_1}{\sigma_0/\sqrt{n}} < -\frac{\mu_1}{\sigma_0/\sqrt{n}} + z_{1-\alpha/2}\right) \\
&= 1 - P\left(\frac{\mu_0 - \mu_1}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} < \frac{\bar{X} - \mu_1}{\sigma_0/\sqrt{n}} < \frac{\mu_0 - \mu_1}{\sigma_0/\sqrt{n}} + z_{1-\alpha/2}\right) \\
&= 1 - P\left(d\sqrt{n} - z_{1-\alpha/2} < Z < d\sqrt{n} + z_{1-\alpha/2}\right)
\end{aligned}
$$

where $Z \sim N(0,1)$ and we defined the **effect size**:

$$d = \frac{\mu_0 - \mu_1}{\sigma_0}$$

Effect Size

Defining the effect size in *this* way is a really smart thing to do.

Many introductions to power calculations (like our textbook) ask you to specify μ_1 and σ separately.

Typically practitioners will plot the power function as a function of $\mu_1 \ldots$ but where do you get σ? Previous experiment? Guess?

Defining the effect size to be the *number of standard deviations that μ_1 is away from μ_0* circumvents this problem while still retaining interpretability.

Z-Test Power

It follows that the power of the Z-test to detect an effect of size d in a sample of size n, rejecting at the α significance level, is

$$\eta(d, n, \alpha) = 1 - (\Phi(d\sqrt{n} + z_{1-\alpha/2}) - \Phi(d\sqrt{n} - z_{1-\alpha/2}))$$

where $\Phi(\cdot)$ is the standard normal CDF.

Z-Test Power

Note the form of the power function: it's an interval of length $2z_{1-\alpha/2}$ under the normal curve, but shifted by $d\sqrt{n}$.

It won't have probability α, unless $d = 0$ (which means H_0 is true).

As either d or n get large, the interval shifts farther and farther away from 0 until it is in an area with effectively no probability- the power goes to 1 as either of these go to ∞.

This means that for fixed effect size, we can make the sample large enough to detect it with as high a probability as we want, and for a fixed sample size, there always exists an effect we can detect with as high a probability as we want.

Finding Power

The power is a function of 3 variables:

- The effect size $d = \frac{\mu_0 - \mu_1}{\sigma}$
- The sample size n
- The significance level α

We have control over all of these in designing our experiment.

Doing a Power Calculation

We get around the unknowns by speaking in terms of effect sizes, as in the derivation.

The effect size d is how many standard deviations the true mean μ_1 is away from our hypothesized mean μ_0.

We say something like "the power of the test to detect a 0.1 SD deviation from μ_0 at the 95% significance level with a sample size of 100 is 90%".

Note: the textbook calls $|\mu_0 - \mu_1|$ the "effect size". I will be clear on tests that I mean my definition.

Doing a Power Calculation

For example, calculate the power with which an effect of $d = 0.1$ can be detected by this test in a sample of size $n = 100$ at the $\alpha = 0.05$ significance level.

Calculate $d\sqrt{n} = 1$, so

$$\begin{aligned} \eta(0.1, 100, 0.05) &= 1 - (\Phi(2.96) - \Phi(-0.96)) \\ &= 1 - (0.9985 - 0.1685) \\ &= 0.17 \end{aligned}$$

That is really small- but $d = 0.1$ is a really small effect.

Doing a Power Calculation

For example, calculate the power with which an effect of $d = 0.5$ can be detected by this test in a sample of size $n = 100$ at the $\alpha = 0.05$ significance level.

Calculate $d\sqrt{n} = 5$, so

$$\begin{aligned}\eta(0.5, 100, 0.05) &= 1 - (\Phi(6.96) - \Phi(3.04)) \\ &= 1 - (1 - 0.9988) \\ &= 1\end{aligned}$$

The power to detect a bigger effect is larger the power to detect a smaller effect, for the same sample size.

Doing a Power Calculation

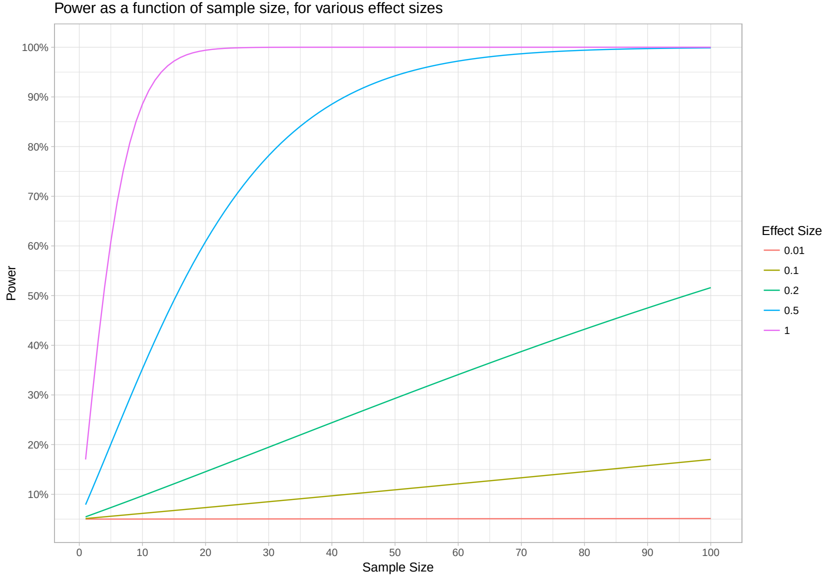
For example, calculate the power with which an effect of $d = 0.1$ can be detected by this test in a sample of size $n = 1,000$ at the $\alpha = 0.05$ significance level.

Calculate $d\sqrt{n} = 3.162$, so

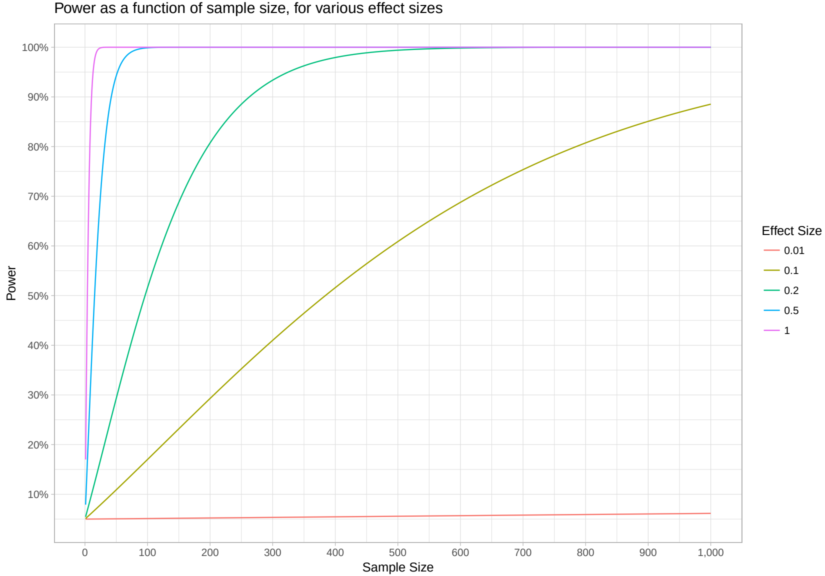
$$\begin{aligned} \eta(0.1, 10000, 0.05) &= 1 - (\Phi(5.12) - \Phi(1.2)) \\ &= 1 - (1 - 0.8853) \\ &= 0.89 \end{aligned}$$

The power to detect the same effect is larger for larger sample sizes.

Power Curve


Power as a function of sample size, for various effect sizes
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
Power
0
10
20
30
40
50
60
70
80
90
100
Sample Size
Effect Size
0.01
0.1
0.2
0.5
1

Power Curve


Power as a function of sample size, for various effect sizes
Power
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0
100
200
300
400
500
600
700
800
900
1,000
Sample Size
Effect Size
0.01
0.1
0.2
0.5
1

Side Note: t-test

Side note: why are we doing this for the Z-test and not the t-test?

Short answer: for a t-test, we plug in the sample standard deviation for the population standard deviation.

But we do a power analysis before seeing the data- so we don't have a sample standard deviation.

There is nothing stopping us from doing the calculations for the t-test, but ironically, the presence of the sample variance in this case precludes us from actually evaluating the power.

We get around this, again, by speaking in terms of the effect size.

Statistical vs Practical Significance

Power and sample size are a crucial aspect of study design, and they are extremely subjective.

Deciding what effect size is of interest in an experiment often requires an enormous amount of domain-specific literature review, and the scientist will often attempt to defer to the statistician (you).

Collaboration is always to be encouraged, but it is important to keep clear the two types of significance: **statistical** and **practical**.

Statistical vs Practical Significance

Loosely speaking, something is **statistically** significant if it is thought to not have happened by chance alone; we have been using this concept when we reject null hypotheses.

We say that if we reject $H_0 : \mu = \mu_0$ at the 5% significance level, then we have observed a *statistically significant* deviation from μ_0 (at this significance level).

Statistical vs Practical Significance

Practically significant means "we care about this, because... science".

For example: an experiment to test whether a new fertilizer results in taller wheat was performed. The experiment showed that the fertilizer resulted in wheat stalks that were 1 inch taller on average, and this result was found to be statistically significant at the 5% significance level.

...but are 1-inch taller wheat stalks something that we care enough about to switch treatments? I don't know, but *this distinction is independent from whether or not the observed effect is concluded to have been from random chance or not*.

Statistical vs Practical Significance

Practically significant means you care that what happened happened.

Statistically significant means that what happened didn't just happen by luck.

You need both to make a reasonable scientific conclusion.

Practical without statistical: "We saw something great but it might not happen if we repeated the experiment".

Statistical without practical: "We saw something completely meaningless, and we might see it again if we repeated the experiment!".

Comparing Tests

When we approached this problem to begin with, we said we wanted to fix α, then find the test with the lowest β (probability of Type II error)- this is equivalent to the test with the highest power.

This suggests that we might compare two tests of the same null hypothesis by evaluating their power.

Comparing Tests

For example, we defined two tests for the mean of a normal distribution when the variance is unknown: the t-test and the likelihood ratio test.

The t-test had rejection region:

$$|t| = \left| \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \right| > t_{1-\alpha/2}$$

The likelihood ratio test had rejection region:

$$-2 \log \Lambda = n \log \left(1 + \frac{t^2}{n-1} \right)$$

Which has higher power?

Comparing Tests

The power of the t-test:

$$\begin{aligned}\eta_T &= 1 - P_T\left(\left|\frac{\bar{X}-\mu_0}{s/\sqrt{n}}\right| > t_{1-\alpha/2}\right)\\ &\approx 1 - P\left(d\sqrt{n} - t_{n-1,1-\alpha/2} < Z < d\sqrt{n} + t_{n-1,1-\alpha/2}\right)\\ &= 1 - (T_{n-1}(d\sqrt{n} + t_{n-1,1-\alpha/2}) - T_{n-1}(d\sqrt{n} - t_{n-1,1-\alpha/2}))\end{aligned}$$

where $T_{n-1}(\cdot)$ is the CDF of a random variable with a t-distribution with $n-1$ degrees of freedom.

Ironically, we have to replace s with σ_0.

Comparing Tests

The power of the LRT is difficult to derive. If $\mu = \mu_1$, then the LRT statistic follows a distribution that we are not going to cover: the *non-central* χ^2 distribution, with 1 degree of freedom and non-centrality parameter

$$\lambda = n\frac{(\mu_0 - \mu_1)^2}{\sigma_0^2} = nd^2$$

The density of the non-central χ^2 involves an infinite series.

You can compute probabilities in `R` using the `ncp` parameter in the `pchisq`, `qchisq`, etc. functions.

LRT Power

The power of the LRT is then

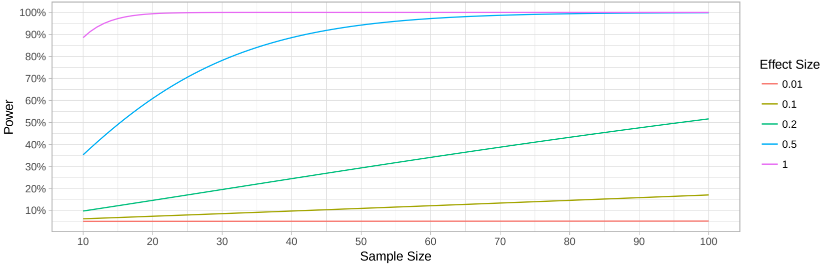
$$\begin{aligned}\eta_{LRT}(d,n,\alpha) &= P\left(\chi^2_1(\lambda = nd^2) > \chi^2_{1,1-\alpha}\right)\\ &= 1 - P_\lambda\left(\chi^2_{1,1-\alpha}\right)\end{aligned}$$

where $P_\lambda(\cdot)$ is the cdf of a non-central χ^2_1 distribution and $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2_1 distribution.

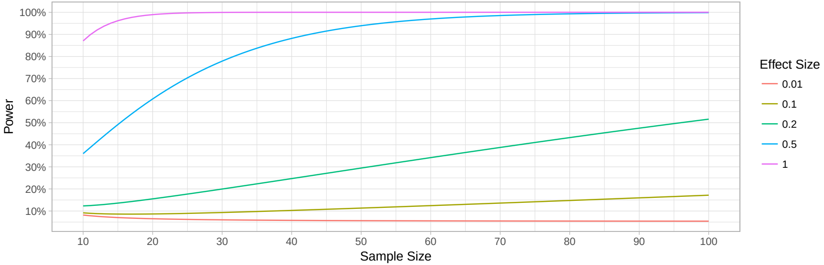
Let's compare the power of these tests graphically.

Comparing Tests

Power of the LRT



Power of the t−test



Compare Tests

They look pretty similar…

d	n	power_lrt	power_t
0.5	10	0.3526081	0.3599515
0.5	20	0.6087795	0.6075203
0.5	30	0.7819080	0.7787822
0.5	50	0.9424375	0.9392206
0.5	100	0.9988173	0.9984867

They are *really* close- note though that the LRT values are only approximations, because evaluating the CDF of the non-central χ^2 requires approximations.

Most Powerful Tests

From a mathematical standpoint, the next logical question is: does there exist a test that is *most powerful* for testing a null against a particular alternative?

Sometimes!

Suppose we are testing $H_0 : \mu = \mu_0$ against the simple alternative $H_1 : \mu = \mu_1$.

We have the following theorem to support us finding a most-powerful test.

Neyman-Pearson Lemma (textbook, page 332)

Theorem: Neyman-Pearson Lemma: for testing the preceding simple null hypothesis against the simple alternative at the α significance level, the Likelihood Ratio Test is most powerful. That is, the test with critical region

$$\frac{f_1(x)}{f_0(x)} > c_\alpha$$

has higher power than any other test in this situation.

Here f_0 and f_1 denote the density of X under H_0 and H_1.

Neyman-Pearson Lemma (textbook, page 332)

Proof: let

$$R_\alpha = \left\{ x : \frac{f_0(x)}{f_1(x)} > c_\alpha \right\}$$

Then by definition of the critical value, under H_0 we have

$$\int_{R_\alpha} f_0(x) dx = \alpha$$

Let R'_α be any other critical region satisfying the above. For *any* density, we can write

$$\int_{R_\alpha} f(x) dx - \int_{R'_\alpha} f(x) dx = \int_{R_\alpha \cap R'^c_\alpha} f(x) dx - \int_{R'_\alpha \cap R^c_\alpha} f(x) dx$$

Neyman-Pearson Lemma (textbook, page 332)

Taking $f = f_0$, we see that the expression is 0 by definition, since we defined R_α and R'_α to have the same probability (α) under H_0.

Taking $f = f_1$, we see that if $x \in R_\alpha^c$ then $c_\alpha f_0(x) > f_1(x)$, where if $x \in R_\alpha$ then $f_1(x) \geq c_\alpha f_0(x)$.

Combining these,

$$\int_{R_\alpha \cap R_\alpha'^c} f_1(x)dx - \int_{R'_\alpha \cap R_\alpha^c} f_1(x)dx \geq \\ c_\alpha \left(\int_{R_\alpha \cap R_\alpha'^c} f_0(x)dx - \int_{R'_\alpha \cap R_\alpha^c} f_0(x)dx \right) = 0$$

Neyman-Pearson Lemma (textbook, page 332)

But the power of the test is the probability of rejecting H_0 when $f = f_1$; that is,

$$\eta_{R_\alpha} = \int_{R_\alpha} f_1(x)dx$$

$$\eta_{R'_\alpha} = \int_{R'_\alpha} f_1(x)dx$$

So we have established our desired result.

Uniformly Most Powerful Tests

The likelihood ratio test is most powerful for testing simple null hypotheses against simple alternatives.

When we test against composite alternatives, a similar notion exists: we say a test is **uniformly most powerful** if it is most powerful against every possible simple alternative.

So if $H_1 : \mu \in \Omega_1$ where Ω_1 is not a singleton set, then a test is said to be UMP if it is most powerful for every possible $\mu \in \Omega_1$.

A UMP test does not always exist.

The likelihood ratio test is UMP in many situations, and it is not UMP typically in situations where no such test exists- so it is the domininant choice.

Some Thoughts on Power

Power calculations and theory are/is really hard.

Usually the distributions involved are intractible and hard to interpret.

There are also philosophical components: tests that are optimal for certain alternatives may be suboptimal for others; care is required to ensure we don't "cheat" and pick our hypotheses (the matter of scientific interest under consideration) based on what tests are optimal.

But, this is done all the time.

We can't avoid power calculations though, because they inform a crucial aspect of study design: sample size determination.

Determining Sample Size

Power calculations are most often done in practice in order to answer the question: how many observations to sample?

Typically n is chosen so that a statement can be made as follows: "we designed our study to detect an effect of size 0.2 at the 95% significance level with a probability of 90%".

In that statement, $\alpha = 0.05$, $\eta = 0.90$, $d = 0.2$- and we would solve the power function for n.

Determining Sample Size

In this example, we can determine n numerically by evaluating the power function:

n	power
10	0.0969354
20	0.1454725
30	0.1947752
50	0.2929889
100	0.5160053
190	0.7872309
195	0.7975459
200	0.8074304

So we need $n = 200$ or so.

Extended Example

Let's look at a more detailed example: in our coin flip example from before, let's design an experiment to be able to make a reasonable conclusion about the fairness of the coin.

We have some decisions to make:

- The significance level, α
- The power with which we would like to reject $H_0 : \theta = 0.5$
- The deviation from H_0 we would like to be able to detect- how unfair does the coin need to be before we conclude that it is unfair?
- The number of times to flip the coin, which will be determined by the above quantities

The Significance Level

When we choose a significance level, we use a mixture of

- Common sense
  - We don't choose α to be absurdly high, such that any H_0 will be rejected
- Best practice as determined by the particular field in which you are working
  - E.g. 0.05 in many life sciences, 0.001 in certain engineering quality control applications, etc.
- Empirical evaluation of the sensitivity of the procedure to this choice
  - Evaluate the tradeoff between sample size, effect size, and α, and make sure that your experiment is robust to at least small changes in α

For our coin flip example, let's choose the usual $\alpha = 0.05$ as is standard in many life sciences.

The Power

We get to choose the power too. We would like to be able to detect a deviation from H_0 with a certain probability.

In practice this means: if the coin really is unfair to the degree that we decide we're interested in, what are the chances that we are going to flip the coin n times, and get a result that allows us to reject H_0?

Determining this is much like determining α: there might be an industry standard like 80% or 90%, but we should also evaluate the tradeoff between power and effect size/sample size.

The Effect Size

This is the deviation from H_0 that we want to be able to detect. How unfair does our coin need to be? $\theta = 0.7$? $\theta = 0.51$?

This is subjective, and again, empirical evaluation can help.

Power Function

In order to do this, we need the power function for the coin tossing experiment. We use our Normal approximation to the binomial that we got using the CLT, so our rejection region is

$$\left|\frac{\bar{X}-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{\sqrt{n}}}}\right| > z_{1-\alpha/2}$$

The only added complication is that now the standard deviation changes between the null and alternative hypotheses.

Power Function

Exercise (assignment 10): show that the power function is

$$\eta = 1 - \Phi\left(d\sqrt{n} + \sqrt{\frac{\theta_0}{\theta_1}\frac{1-\theta_0}{1-\theta_1}}z_{1-\alpha/2}\right) + \Phi\left(d\sqrt{n} - \sqrt{\frac{\theta_0}{\theta_1}\frac{1-\theta_0}{1-\theta_1}}z_{1-\alpha/2}\right)$$

where the effect size in this problem is

$$d = \frac{\theta_0 - \theta_1}{\sqrt{\theta_1(1-\theta_1)}}$$

Designing the Experiment

Let's do some empirical study to determine the quantities we need.

First: what were we doing before? We threw the coin 10 times. What effect size/power tradeoff did that give?

n	theta1	power
10	0.55	0.0603442
10	0.60	0.0918014
10	0.70	0.2243332
10	0.80	0.4688166
10	0.90	0.8288838
10	0.99	1.0000000

Designing the Experiment

It turns out our experiment with $n = 10$ was very low power. The coin would have to been very unfair ($\theta = 0.90$) for us to have any reasonable chance of detecting this in 10 flips.

We can examine how increasing the number of flips changes this:

n	theta1	power
10	0.55	0.060
20	0.55	0.072
50	0.55	0.108
100	0.55	0.169
500	0.55	0.609

Designing the Experiment

n	theta1	power
10	0.6	0.092
20	0.6	0.140
50	0.6	0.289
100	0.6	0.516
500	0.6	0.995
10	0.7	0.224
20	0.7	0.426
50	0.7	0.828
100	0.7	0.987
500	0.7	1.000

Designing the Experiment

For example, if we wanted to detect a coin having a 70% chance of heads with about 82% probability, we'd need to throw it 50 times.

If we wanted to detect a coin having a 60% chance of heads with about 99% probability, we'd need to throw it 500 times.

So on, and so forth.

We could also increase our significance level α to increase power-but we need to be very careful. Are we willing to increase the probability of Type I Error in order to have a greater chance of detecting a false null?

Are we willing to risk sending more innocent people to prison, to increase our chances of convicting a guilty person?