Hypothesis Testing

Rebecca Kurtz-Garcia

University of California Riverside, Grad Quant

October 23, 2019

Why Hypothesis testing?

You already know how to estimate the parameter θ.

Now you want to know if the value of the parameter θ is equal to θ_0, and quantify how confident you are.

A wonderful thing about hypothesis testing is that we recognize that our results and conclusion may be wrong and we quantify by how much.

A Note On Notation

- For population parameters we use Greek letters (μ, σ, ρ, θ, etc)

- For sample estimates of the population parameters we use hats or letters ($\widehat{\mu}, \widehat{\theta}, \overline{x}, s$, etc)

What can we use hypothesis testing for?

To infer things about the population, such as...

- parameters (mean, variance, proportions, etc.)
- operations of parameters (differences, ratios, etc.)
- comparing parameters across multiple populations
- MORE!

Where do Hypothesis Tests Come From

- The basic thought process is least 200 years old, possibly as early as 1660s.
- Happy accident? Karl Pearson wanted to analyze biological data, developed a more formal hypothesis testing procedure and the -Squared distribution in the process. (There is a reason -Squared distribution comes up so much, it was literally designed for this!)
- 1920 - 1930s; R.A. Fisher published papers expanding the idea with different distributions. Made the idea very versatile.
- 1938; G.W. Snedecor front the first textbook, which popularized tests. It was at one point the most cited scientific book or paper of all time.
- Late 1920s; E. Pearson and J. Neyman revamped the theory. Neyman-Pearson Lemma, proof of having the "Best Test". $\leftarrow$ This is what we mostly use today in introductory courses.
- E. Lehmann expanded the idea even more, making it more versatile and published a textbook.

Basic Steps In Hypothesis Testing

1. Make Assumption
2. Collect/acquire data related to assumption
3. Reject assumption or not based on data.
    - P-value***
    - Critical Value
    - Confidence Interval

Basic Steps in Hypothesis Testing (version 2)

1. Formulate the null hypothesis H_0 and H_A
2. Identify the distribution of the sample estimate when H_0 is true.
3. Select a significance level, α (usually 0.1, 0.05, 0.01).
4. Compute the p-value. The p-value is the probability that we observe our data or something more extreme given the H_0 is true.
5. Compare the p-value to a significance level. Determine if H_0 should be rejected (p-value $\leq \alpha$) or if we fail to reject it (p-value $> \alpha$).

How to format H_0 and H_A

We may consider the following structures:

- $H_0 : \theta = \theta_0 \qquad H_A : \theta \neq \theta_0$
- $H_0 : \theta \leq \theta_0 \qquad H_A : \theta > \theta_0$
- $H_0 : \theta \geq \theta_0 \qquad H_A : \theta < \theta_0$
- $H_0 : \theta = \theta_0 \qquad H_A : \theta = \theta_A, \theta_0 \neq \theta_A$

Note that the H_0 always has an equals sign. This is critical for the construction of the tests.

H_0 is called the Null-Hypothesis because most hypothesis testing is focused on lack of an effect.

Example

An alien empire is considering taking over planet Earth, but they will only do so if the portion of rebellious humans is less than 10%, percent. They abduct a random sample of 400 humans, performed special psychological tests, and found that 14%, percent of the sample are rebellious. The population parameter is $\theta=$ percent of rebellious humans.

What is H_0 vs H_A for the alien's test?

Example

An alien empire is considering taking over planet Earth, but they will only do so if the portion of rebellious humans is less than 10%, percent. They abduct a random sample of 400 humans, performed special psychological tests, and found that 14%, percent of the sample are rebellious. The population parameter is $\theta=$ percent of rebellious humans.

What is H_0 vs H_A for the alien's test?

$H_0 : \theta \leq .10$

$H_A : \theta > .10$

Suppose $\alpha = 0.05$, $n = 400$, $\widehat{\theta} = 0.14$, and p-value $= 0.01$.

What is the conclusion of the test?

Example

An alien empire is considering taking over planet Earth, but they will only do so if the portion of rebellious humans is less than 10%, percent. They abduct a random sample of 400 humans, performed special psychological tests, and found that 14%, percent of the sample are rebellious. The population parameter is $\theta =$ percent of rebellious humans.

What is H_0 vs H_A for the alien's test?

$H_0 : \theta \leq .10$

$H_A : \theta > .10$

Suppose $\alpha = 0.05$, $n = 400$, $\widehat{\theta} = 0.14$, and p-value $= 0.01$.

What is the conclusion of the test?

p-value $\leq \alpha \rightarrow$ reject H_0. Humans are too rebellious.

Example

An alien empire is considering taking over planet Earth, but they will only do so if the portion of rebellious humans is less than 10%, percent. They abduct a random sample of 400 humans, performed special psychological tests, and found that 14%, percent of the sample are rebellious. The population parameter is $\theta =$ percent of rebellious humans.

What is H_0 vs H_A for the alien's test?

$H_0 : \theta \leq .10$

$H_A : \theta > .10$

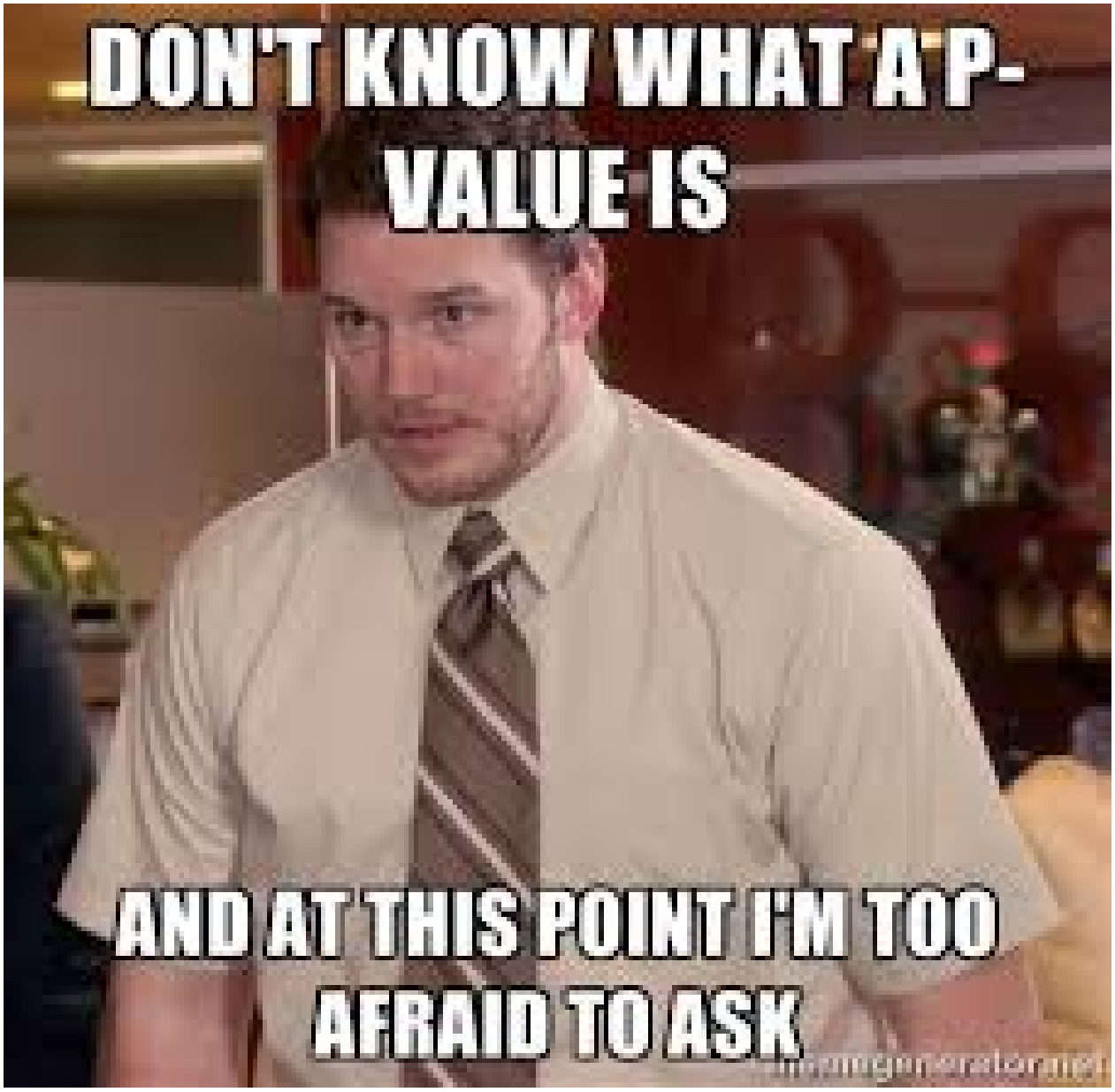
Suppose $\alpha = 0.05$, $n = 400$, $\widehat{\theta} = 0.14$, and p-value $= 0.01$.

What is the conclusion of the test?

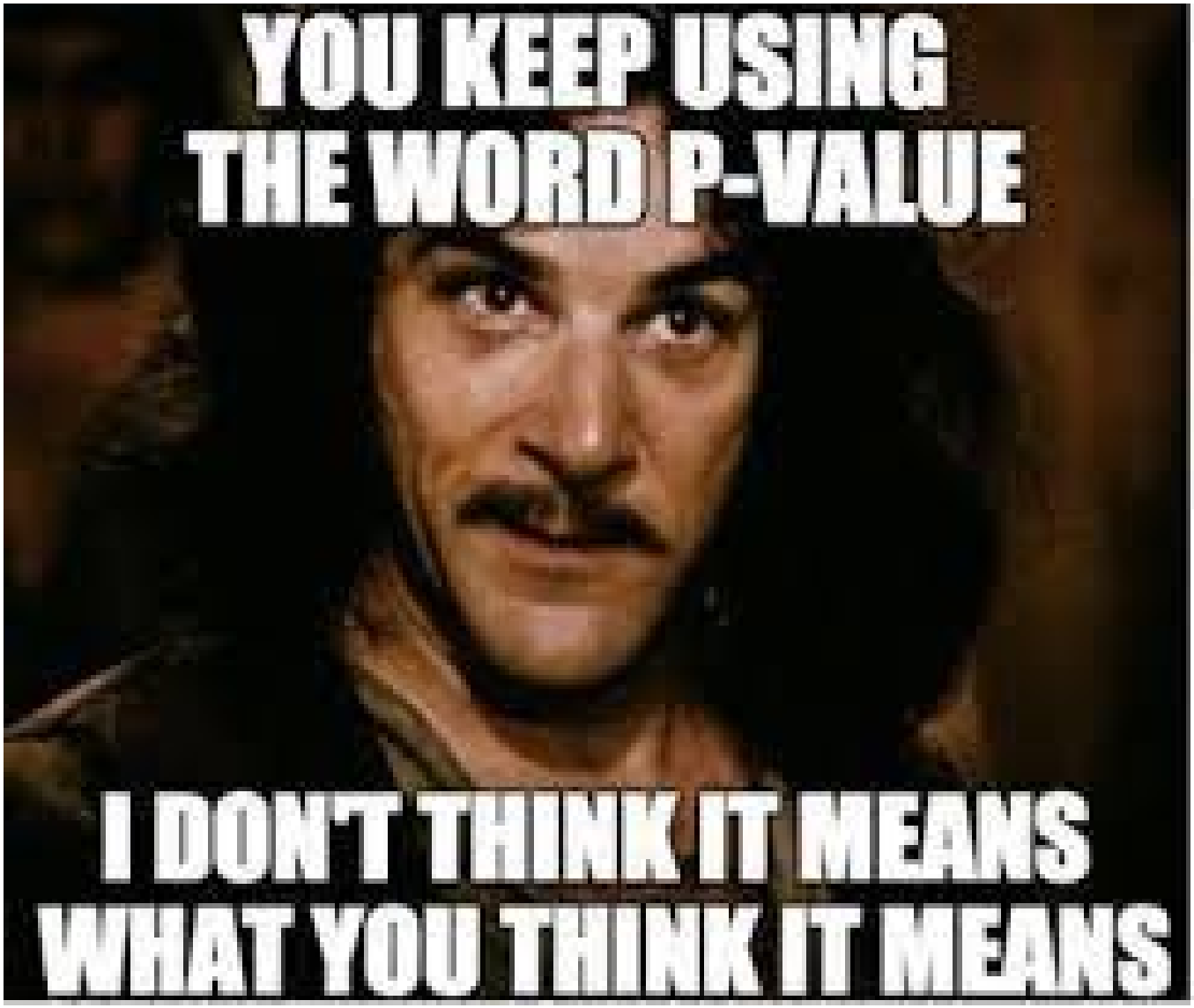
p-value $\leq \alpha \rightarrow$ reject H_0. Humans are too rebellious.

Why do we use the p-value in this way?

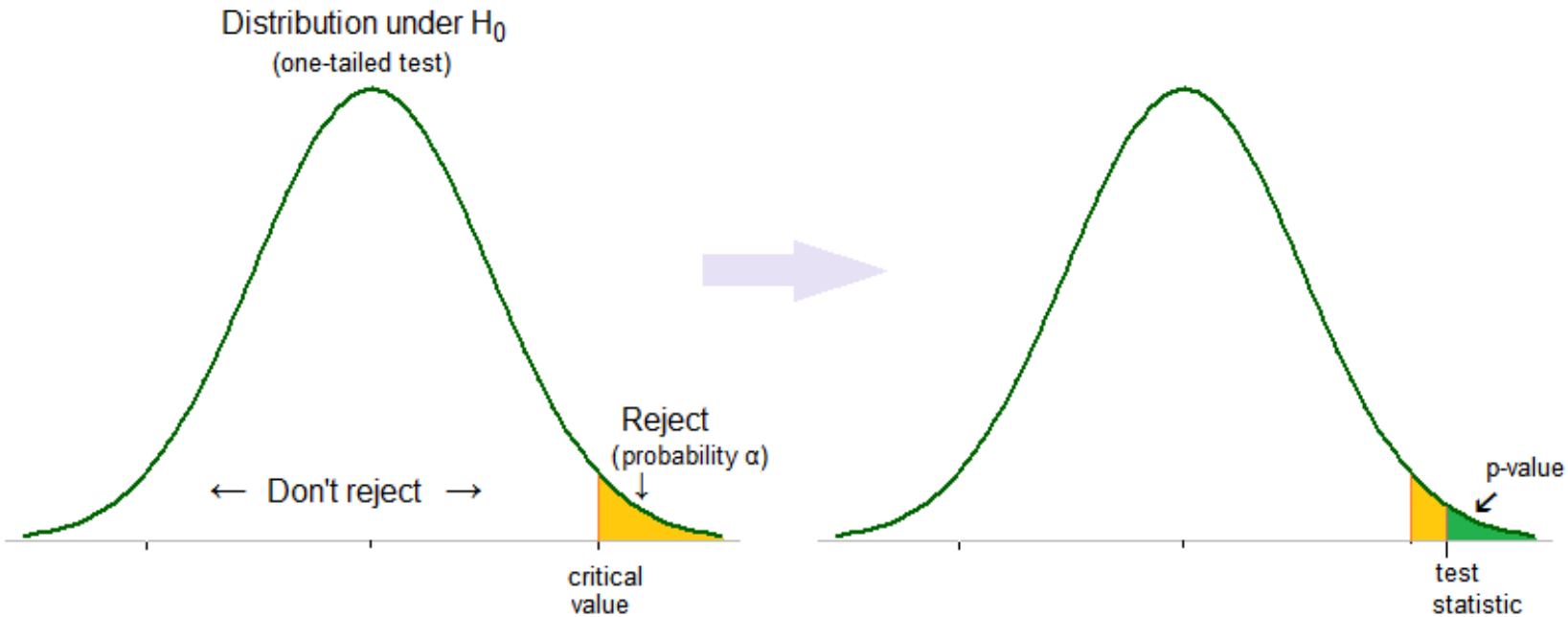
What's in a P-Value anyways?



What's in a P-Value anyways?



What's in a P-Value anyways?



Recall: Definition of a p-value is observing our data or something more extreme given the null hypothesis is true.

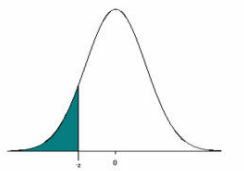
It is still reasonable to observe data in the area of α, it is just not likely.

How do find the p-value?

Use a test statistic, find corresponding p-value for your test statistic.

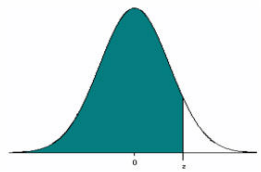
There are different charts for different distributions.

Table of Standard Normal Probabilities for Negative Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table of Standard Normal Probabilities for Positive Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

You can also calculate the p-value directly...

Calculating directly is hard. This is why we use charts alot.

$$P(X > x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Example

A certain fertilizer company makes bags of fertilizer, it is believed that the weight of these bags is normally distributed with a mean of 7.4kg and a standard deviation of 0.15kg. They take a sample of these bags from the assembly lines to test if the mean fertilizer weight is not 7.4.

What is H_0 vs H_A?

Example

A certain fertilizer company makes bags of fertilizer, it is believed that the weight of these bags is normally distributed with a mean of 7.4kg and a standard deviation of 0.15kg. They take a sample of these bags from the assembly lines to test if the mean fertilizer weight is not 7.4.

What is H_0 vs H_A?

$H_0 : \mu = 7.4$

$H_A : \mu \neq 7.4$

Suppose $\alpha = 0.05$, $n = 50$, $\overline{x} = 7.36$kg, $s = .12$kg, and p-value $= 0.01$.

What do we conclude?

Example

A certain fertilizer company makes bags of fertilizer, it is believed that the weight of these bags is normally distributed with a mean of 7.4kg and a standard deviation of 0.15kg. They take a sample of these bags from the assembly lines to test if the mean fertilizer weight is not 7.4.

What is H_0 vs H_A?

$H_0 : \mu = 7.4$

$H_A : \mu \neq 7.4$

Suppose $\alpha = 0.05$, $n = 50$, $\overline{x} = 7.36$kg, $s = .12$kg, and p-value $= 0.01$.

What do we conclude?

If the p-value $\leq \frac{\alpha}{2}$. Reject H_0, the mean weight is not 7.4kg.

Example

A certain fertilizer company makes bags of fertilizer, it is believed that the weight of these bags is normally distributed with a mean of 7.4kg and a standard deviation of 0.15kg. They take a sample of these bags from the assembly lines to test if the mean fertilizer weight is not 7.4.

What is H_0 vs H_A?

$H_0 : \mu = 7.4$

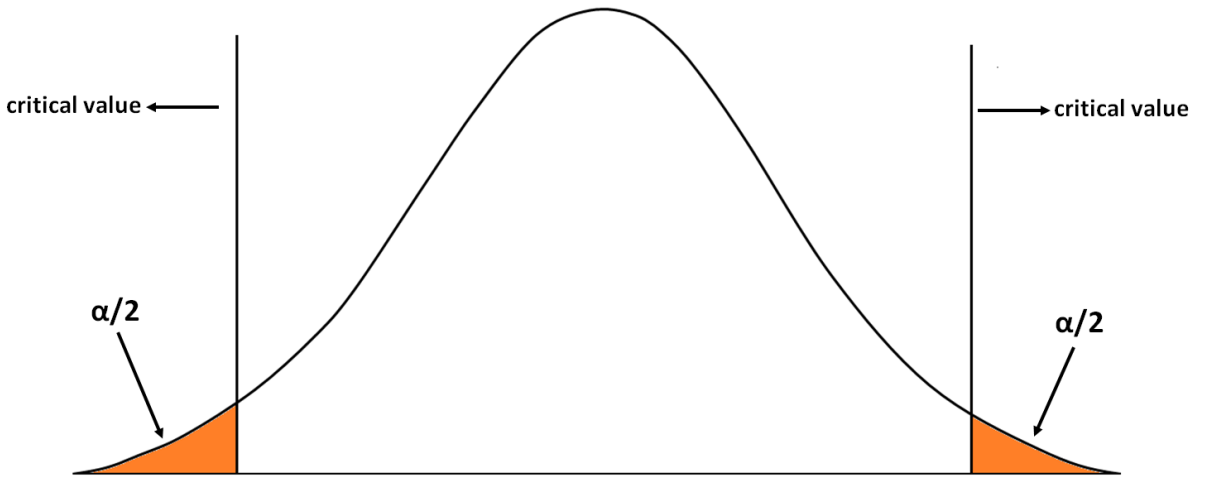
$H_A : \mu \neq 7.4$

Suppose $\alpha = 0.05$, $n = 50$, $\overline{x} = 7.36$kg, $s = .12$kg, and p-value $= 0.01$.

What do we conclude?

If the p-value $\leq \frac{\alpha}{2}$. Reject H_0, the mean weight is not 7.4kg.

Example



What do we get from Hypothesis Testing

We learn if it would be reasonably plausible to *infer* H_0 in true. (Hint: Hypothesis testing is part of inferential statistics)

We do NOT learn:

1. If H_0 is true or false.
2. If H_A is true or false.

We infer, we may never know the truth.

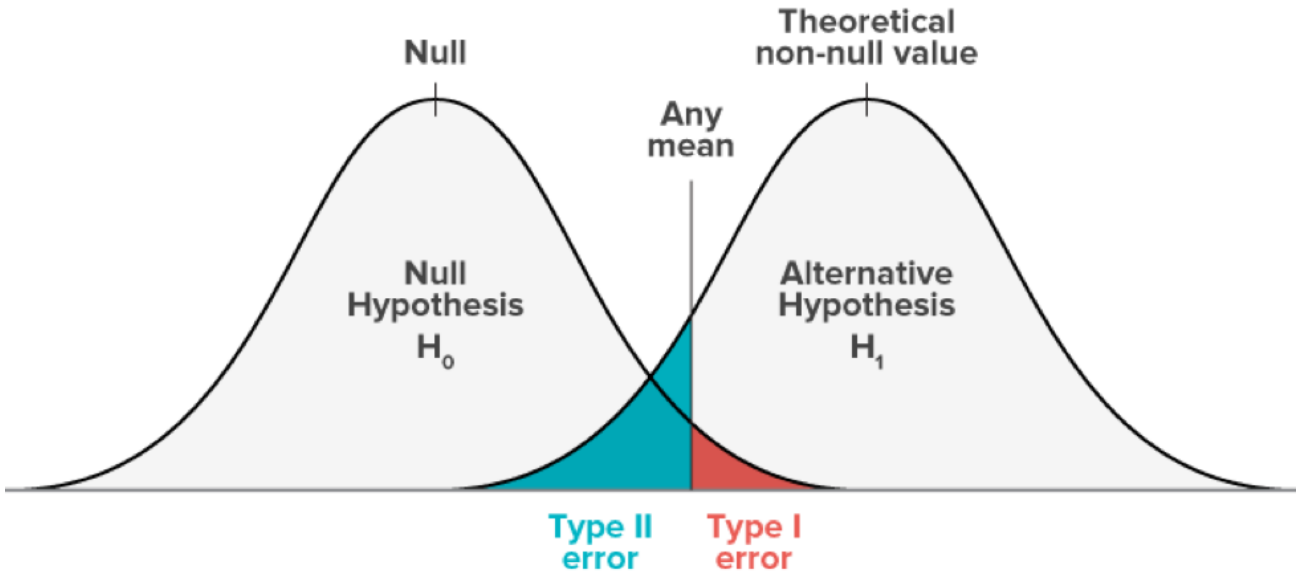
Notice, in all these examples we need to predetermine α. It is also still reasonable to observe data in the area of α, it is just not likely. Always report significance level, α, with your conclusion.

Hypothesis Testing is Not Perfect - Errors Happen

How do we get α? We must balance out the errors.

		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (Power) (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

Hypothesis Testing is Not Perfect- Errors Happen



- A **Type I** error occurs if we reject the null hypothesis H_0 (in favor of the alternative hypothesis HA) when the null hypothesis H_0 is true. We denote $\alpha = \text{P(Type I Error)}$.
- A **Type II** error occurs if we fail to reject the null hypothesis H_0 when the alternative hypothesis H_A is true. We denote $= \text{P(Type II Error)}$

How do we balance the errors?

In practice, we predetermine α. This is based on the needs of the project, and what type of errors you prefer. Then we minimize β or, equivalently maximize power $(1 - \beta)$. This is a balancing act.

The most common α values are 0.05, or 0.10. The β value is typically around 0.20. There is no statistical reasoning for this. The value of α and β is determined by researcher depending on the context of the question.

What size α value would you give for the following scenarios?

How should we balance the errors? (Example 1)

Seth is starting his own food truck business, and he's choosing cities where he'll run his business. He wants to survey residents and test whether or not the demand is high enough to support his business before he applies for the necessary permits to operate in a given city. He'll only choose a city if there's strong evidence that the demand there is high enough.

- H_0:The demand is not high enough
- H_A:The demand is high enough.

Type 1 error:

How should we balance the errors? (Example 1)

Seth is starting his own food truck business, and he's choosing cities where he'll run his business. He wants to survey residents and test whether or not the demand is high enough to support his business before he applies for the necessary permits to operate in a given city. He'll only choose a city if there's strong evidence that the demand there is high enough.

- H_0:The demand is not high enough
- H_A:The demand is high enough.

Type 1 error: He chooses a city where demand isn't actually high enough.

Type 2 error:

How should we balance the errors? (Example 1)

Seth is starting his own food truck business, and he's choosing cities where he'll run his business. He wants to survey residents and test whether or not the demand is high enough to support his business before he applies for the necessary permits to operate in a given city. He'll only choose a city if there's strong evidence that the demand there is high enough.

- H_0:The demand is not high enough
- H_A:The demand is high enough.

Type 1 error: He chooses a city where demand isn't actually high enough.

Type 2 error: He doesn't choose a city where demand is actually high enough.

Which error is more concerning?

How should we balance the errors? (Example 1)

Seth is starting his own food truck business, and he's choosing cities where he'll run his business. He wants to survey residents and test whether or not the demand is high enough to support his business before he applies for the necessary permits to operate in a given city. He'll only choose a city if there's strong evidence that the demand there is high enough.

- H_0:The demand is not high enough
- H_A:The demand is high enough.

Type 1 error: He chooses a city where demand isn't actually high enough.

Type 2 error: He doesn't choose a city where demand is actually high enough.

Which error is more concerning? Type 1. Make α small.

How do we determine the critical value? (Example 2)

Employees at a health club do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment.

- H_0:The water quality is acceptable
- H_A:The water quality is not acceptable.

Type 1 error:

How do we determine the critical value? (Example 2)

Employees at a health club do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment.

- H_0:The water quality is acceptable
- H_A:The water quality is not acceptable.

Type 1 error: The club closes the pool when it doesn't need to be closed.
Type 2 error:

How do we determine the critical value? (Example 2)

Employees at a health club do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment.

- H_0:The water quality is acceptable
- H_A:The water quality is not acceptable.

Type 1 error: The club closes the pool when it doesn't need to be closed.
Type 2 error: The club doesn't close the pool when it needs to be closed.

Which is more dangerous?

How do we determine the critical value? (Example 2)

Employees at a health club do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment.

- H_0:The water quality is acceptable
- H_A:The water quality is not acceptable.

Type 1 error: The club closes the pool when it doesn't need to be closed.
Type 2 error: The club doesn't close the pool when it needs to be closed.

Which is more dangerous? Type 2 Error. Let α be larger.

Now we set α, what about β?

Calculating β is a bit more complicated. We start with a simple example.

How to calculate a Type II error?

Suppose you are interested in the mean of a normal random variable with a known standard deviation, $\sigma = 1$, $n = 25$. It is known the mean of the normal random variable is either 0 or 4.

$$H_0 : \mu = 0$$

$$H_A : \mu = 4$$

How to calculate a Type II error?

Suppose you are interested in the mean of a normal random variable with a known standard deviation, $\sigma = 1$, $n = 25$. It is known the mean of the normal random variable is either 0 or 4.

$$H_0 : \mu = 0$$

$$H_A : \mu = 4$$

Suppose the researcher set $\alpha = 0.05$. Thus the critical value is 1.96.

$$P(\text{Type II error})$$

$$= P(\text{Accept } H_0 | H_A \text{ is true})$$

$$= P(\overline{x} < 1.96 | \mu = 4)$$

Convert the H_A distribution to standard normal.

$$= P(Z < \frac{1.96 - 4}{1})$$

$$= P(Z < -2.04) = 0.0207$$

What if the alternative hypothesis is not simple?

Let X denote the IQ of a randomly selected adult American. Assume, a bit unrealistically, that X is normally distributed with unknown mean μ and standard deviation 16. Take a random sample of n $= 16$ students, so that, after setting the probability of committing a Type I error at $\alpha = 0.05$. We want to test

$H_0 : \mu = 100$

$H_0 : \mu > 100$

What if the alternative hypothesis is not simple?

Let X denote the IQ of a randomly selected adult American. Assume, a bit unrealistically, that X is normally distributed with unknown mean μ and standard deviation 16. Take a random sample of n $= 16$ students, so that, after setting the probability of committing a Type I error at $\alpha = 0.05$. We want to test

$H_0 : \mu = 100$

$H_0 : \mu > 100$

We need to know the true distribution of our parameter to calculate β.

The further the true mean is from the μ under H_0, the smaller β.

Power

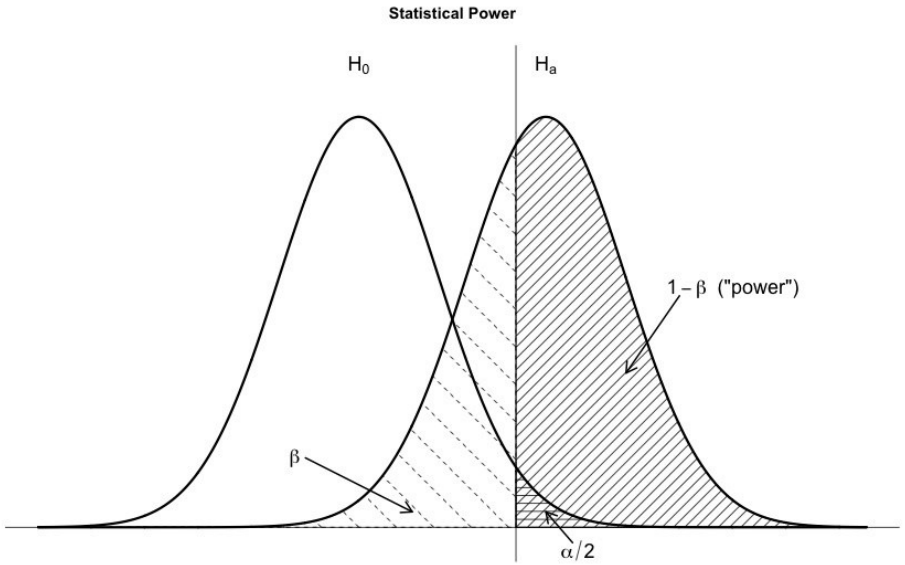
The **power** of a hypothesis test is the probability of correctly rejecting H_0.

That is, the power of a hypothesis test is the probability of rejecting the null hypothesis H_0 when H_0 is not true.

Typically we think in terms of power, instead of Type II errors. That is we want to minimize α and maximize $(1 - \beta)$. Instead of thinking of minimizing α and β.

For the previous example, power $=$ 1 - 0.0207 $=$ .9793.

Balancing Errors



In general, we'll want to do the following:

(1) Minimize the probability of committing a Type I error. That, is minimize α P(Type I Error). Typically, a significance level of $\alpha \leq 0.10$ is desired.

(2) Maximize the power (at a value of the parameter under the alternative hypothesis that is scientifically meaningful). Typically, we desire power to be 0.80 or greater.

Power Function

$$\text{Power } = K(\theta) = 1 - \Phi(Z_\alpha)$$

Where

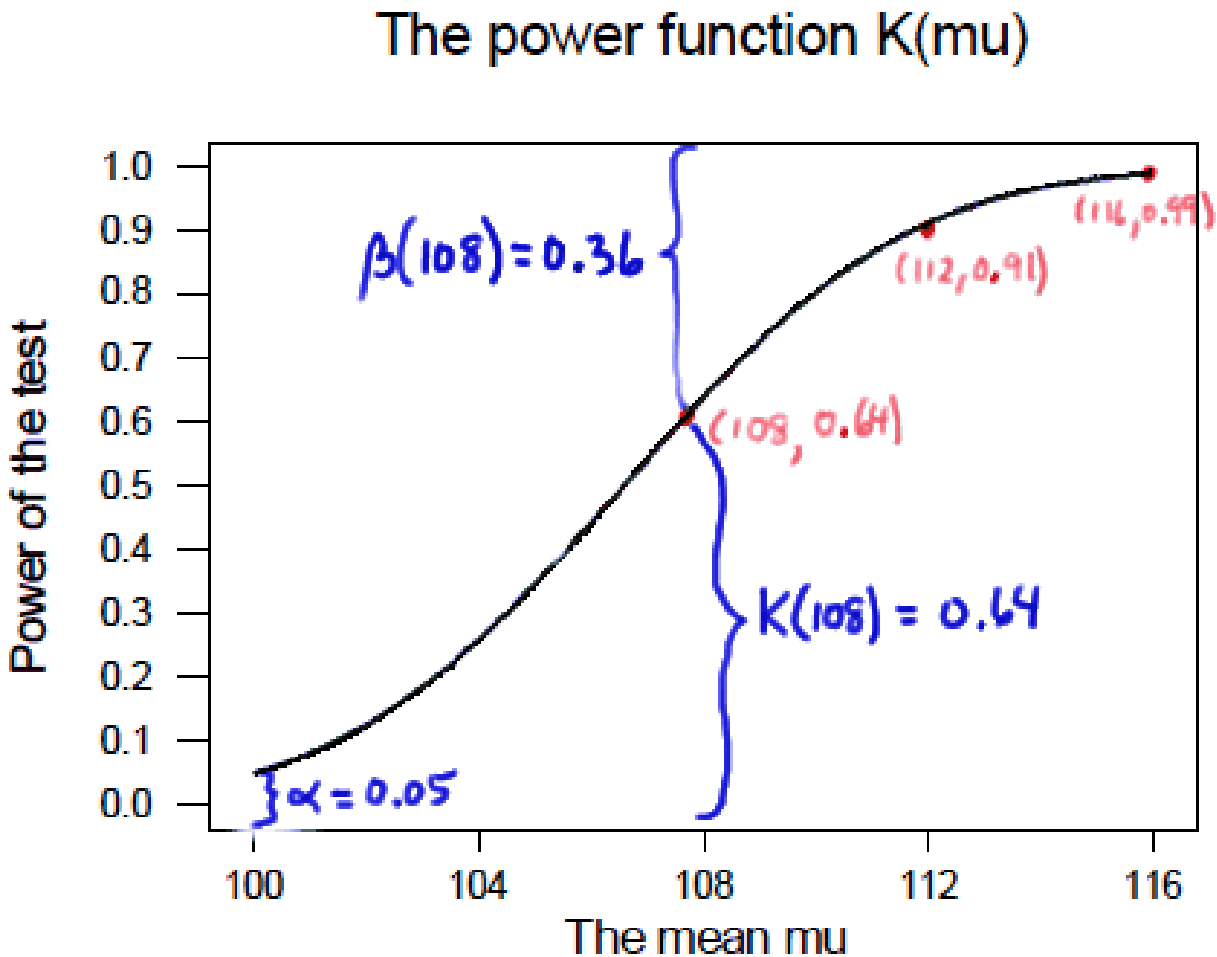
- $\Phi()$ is the cumulative distribution function of the normal distribution
- $Z_\alpha = \frac{\overline{x}_\alpha - \mu}{\sigma/\sqrt{n}}$
- $\overline{x}_\alpha$ is the non standardized critical value under H_0
- μ is the unknown true mean

The power function is a function of the distribution of the true unknown mean. We may never know the true power, but we can still manipulate it.

The power function changes depending on the question at hand.

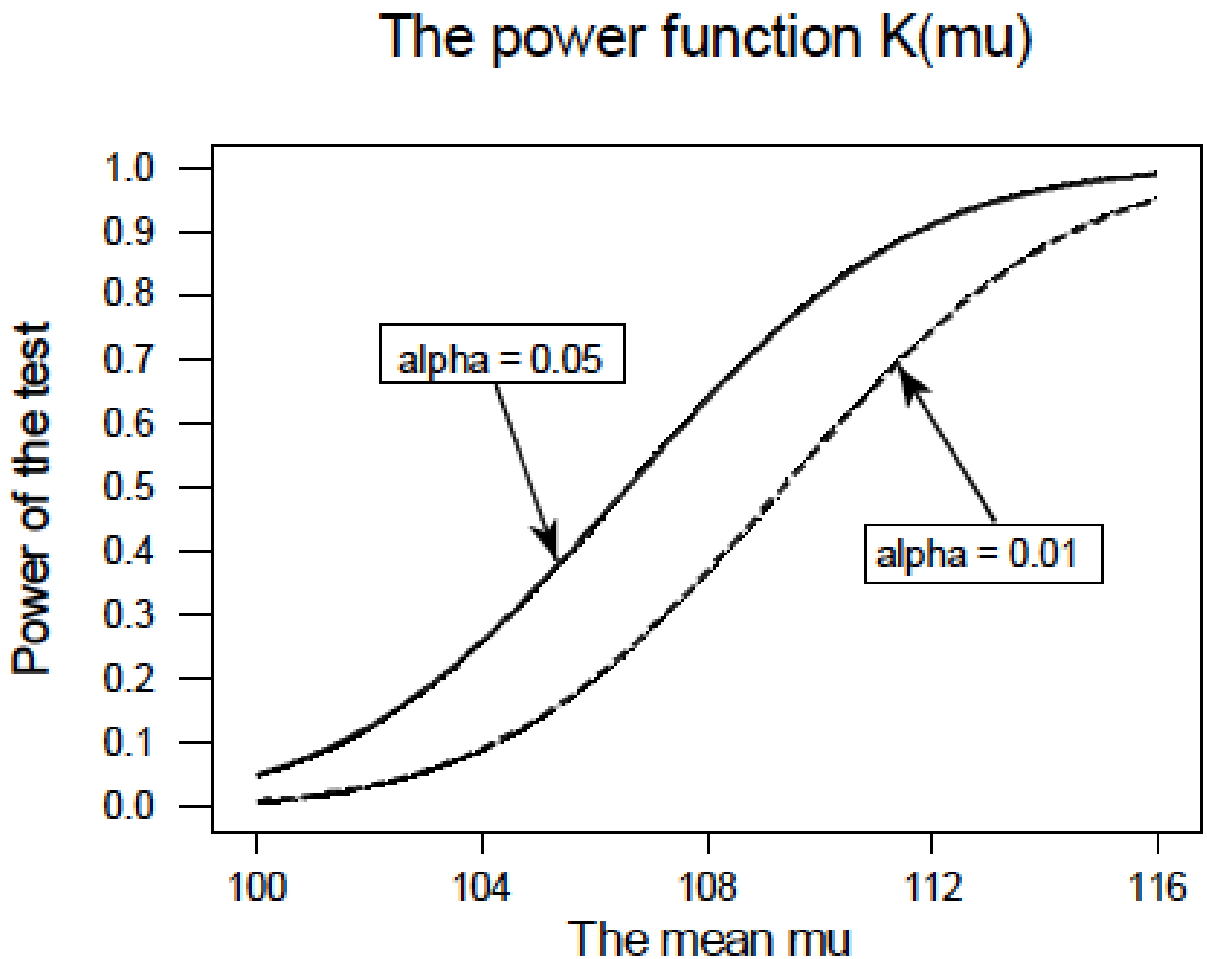
Power Function

Instead of looking at power as a single value, we look at the power function. We can increase power (decrease the probability of a Type 2 error) by increasing α or by increasing the sample size.



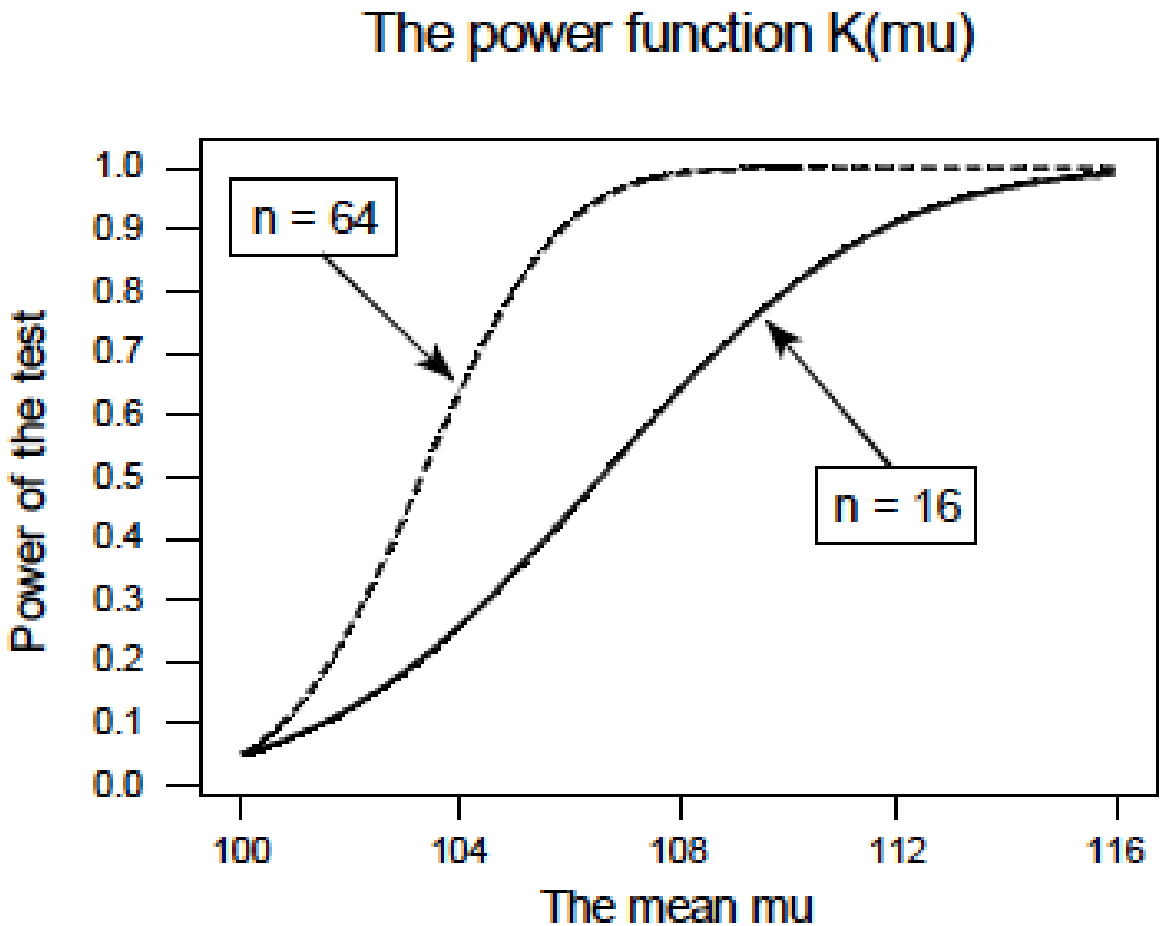
Power Function

Instead of looking at power as a single value, we look at the power function. We can increase power (decrease the probability of a Type 2 error) by increasing α or by increasing the sample size.



Power Function

Instead of looking at power as a single value, we look at the power function. We can increase power (decrease the probability of a Type 2 error) by increasing α or by increasing the sample size.



What other things come up with hypothesis testing?

1. Confidence Intervals
2. Critical Values
3. Testing different parameters
4. Testing multiple populations at once
5. P-Hacking
6. Multiple Hypothesis Testing

Confidence Interval Example

Suppose the average price of a gallon of milk in California is \$2.50. You want to see if the average price of milk in Riverside (μ_R)is the same as the average price in California. You obtain a random sample (n=36) of milk prices in Riverside. Let $\alpha = 0.05$, and $\sigma = 4$ for California and Riverside milk prices.

$H_0 : \mu_R = 2.5$

$H_A : \mu_R \neq 2.5$

Suppose your sample average is $\widehat{\mu}_R = 2.9$ Calculate the confidence interval:

$$\widehat{\mu}_R \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$2.9 \pm 1.96\frac{4}{6}$$

$$(1.593, 4.207)$$

Fail to Reject H_0, interval does not contain hypothesized value.

Difference in Means

A new iPhone came out recently and the advertisements claim that the battery lasts 3 days longer than the previous model. You obtain a random sample of 15 people with the new iPhone, and 20 people with the old iPhone and ask them about their battery life. Test the advertisement claim.

$H_0 : \mu_{new} - \mu_{old} = 3$

$H_A : \mu_{new} - \mu_{old} \neq 3$

Notice:

Difference in Means

A new iPhone came out recently and the advertisements claim that the battery lasts 3 days longer than the previous model. You obtain a random sample of 15 people with the new iPhone, and 20 people with the old iPhone and ask them about their battery life. Test the advertisement claim.

$H_0 : \mu_{new} - \mu_{old} = 3$

$H_A : \mu_{new} - \mu_{old} \neq 3$

Notice:

- small sample size
- populations are independent

Difference in Means

A new iPhone came out recently and the advertisements claim that the battery lasts 3 days longer than the previous model. You obtain a random sample of 15 people with the new iPhone, and 20 people with the old iPhone and ask them about their battery life. Test the advertisement claim.

$H_0 : \mu_{new} - \mu_{old} = 3$

$H_A : \mu_{new} - \mu_{old} \neq 3$

Test statistic: $t^* = \frac{\widehat{\mu}_{new} - \widehat{\mu}_{old} - 3}{SE}$

Where $SE = \sqrt{\frac{\widehat{\sigma}_{new}}{15} + \frac{\widehat{\sigma}_{old}}{20}}$

Pros and Cons

PROS

- Streamline. Step-by-step process.
- Versatile. A lot of expansion on the initial idea.
- Fill-In-The-Blank.
- Theory backs up the method.

CONS

- A major critic was R.A. Fisher himself. Viewed Hypothesis testing as a rough tool that should never stand alone.
- Scientists were left in the dark as to how to choose a significance value α.
- P-values do not have an intuitive means
- Different schools of thought: Bayesian and Likelihood approach to testing theories.
- What constitutes as optimum is arbitrary.

Summary

- Hypothesis Testing is formulaic, but we should still be cautious.
- Hypothesis Testing is versatile.
- Use caution when determining α, consider which errors you like
- Increasing sample size often helps us decrease our Type 2 error rate, without increasing our Type 1 error rate.

Resources

https://newonlinecourses.science.psu.edu/stat414/node/306/

https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat05865

https://stats.stackexchange.com/questions/124178/why-do-we-compare-p-value-to-significance-level-in-hypothesis-testing-of-mean

https://www.abtasty.com/blog/type-1-and-type-2-errors/

https://towardsdatascience.com/hypothesis-testing-using-t-test-inferential-statistics-part3-6fb43683bc32

https://www.quora.com/What-is-a-good-explanation-of-statistical-power-effective-size-and-their-relationship-in-hypothesis-testing-And-how-to-calculate-them

https://sixsigmastudyguide.com/z-scores-z-table-z-transformations/