15.0 More Hypothesis Testing

- Answer Questions
- Type I and Type II Error
- Power Calculation
- Bayesian Hypothesis Testing

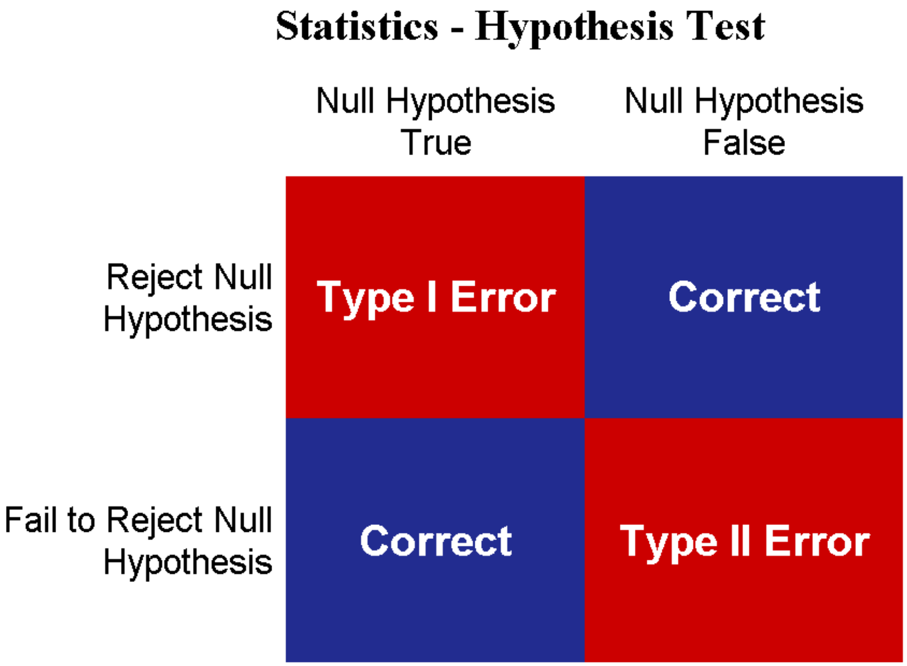
15.1 Type I and Type II Error

In the philosophy of hypothesis testing, the null hypothesis is innocent until proven guilty. You require evidence, from your data, in order to decide against the null hypothesis.

Before you collect your data, you decide on some small probability α (usually 0.05 or 0.01) that will be your threshold for rejecting the null. If your significance probability turns out to be less than that value, then you reject the null hypothesis.

Otherwise, you fail to reject the null hypothesis. Speaking formally, one never "accepts" or "proves" the null hypothesis; one simply fails to reject the null hypothesis.

There are four possible situations:



Statistics - Hypothesis Test

	Null Hypothesis True	Null Hypothesis False
Reject Null Hypothesis	Type I Error	Correct
Fail to Reject Null Hypothesis	Correct	Type II Error

In two of the situations, your test reaches the correct conclusion. But you make a **Type I error** if you reject the null hypothesis when the null hypothesis is true, and you make a **Type II error** if you fail to reject the null when the null hypothesis is false.

Any two of the following three quantities determines the third:

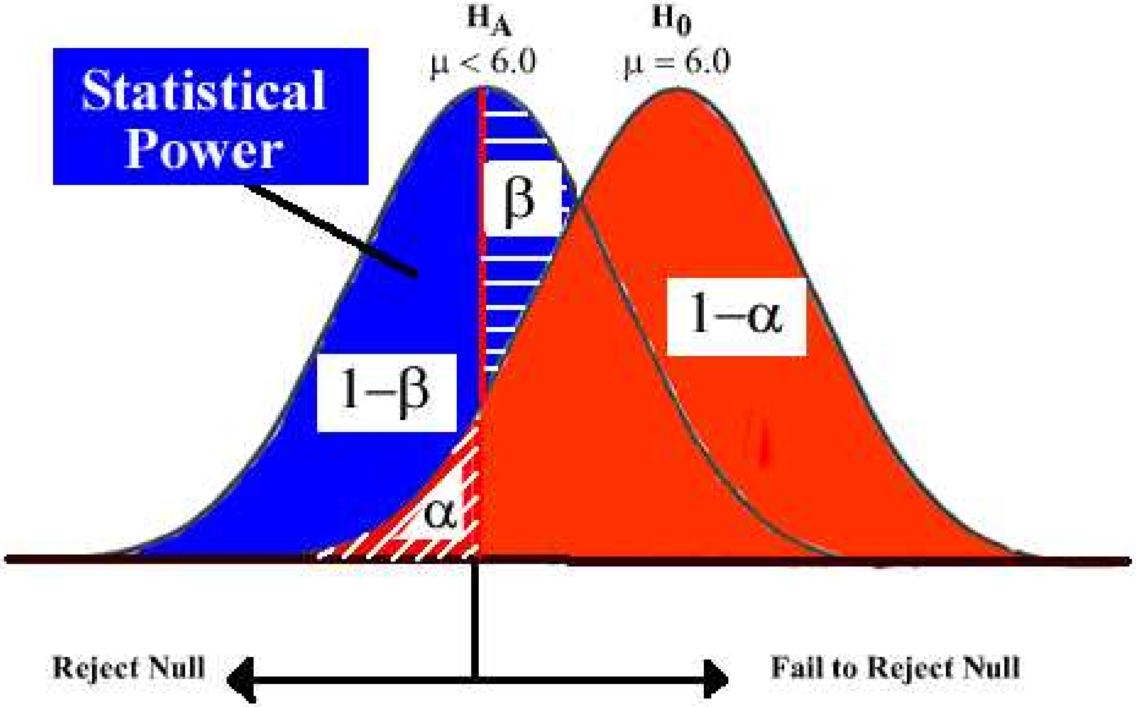
- n, the sample size in the test;
- α, the probability of Type I error; and
- β, which is the probability of Type II error.

Typically, circumstances force you to pick α and n.

The **power** of a test is the $1 - \beta$, which is the probability that your test correctly rejects the null hypothesis when the null hypothesis is false.

In practice, one picks α at the outset, and then obtains the largest sample size n that one can afford, in order to maximize the power of the test.

This figure illustrates the definitions. It assumes a one-sided test of $\mathrm{H}_0: \ \mu \geq 6$ versus $\mathrm{H_A}: \ \mu < 6$ with σ known and some level α (say 0.05).



15.2 The Power of a Test

In many cases one can calculate the power of a test. This is important when deciding how large a sample you need—if your test is underpowered, you can improve it by investing in a larger sample size.

Example 1: You have a sample of size 100 from a normal population with known standard deviation 4. You want to test $H_o : \mu \geq 6$ versus $H_A : \mu < 6$ with a Type I error rate of 0.05.

Suppose the population actually has a true mean of 5. What will be the power of your test?

$$\begin{aligned}
\textbf{power} = 1-\beta &= 1-\mathbb{P}[ts > -1.645] \\
&= \mathbb{P}[\frac{\bar{X}-6}{4/\sqrt{100}} < -1.645] \\
&= \mathbb{P}[\frac{\bar{X}-6}{0.4} < -1.645] \\
&= \mathbb{P}[\frac{\bar{X}-5+5-6}{0.4} < -1.645] \\
&= \mathbb{P}[\frac{\bar{X}-5}{0.4} + \frac{5-6}{0.4} < -1.645] \\
&= \mathbb{P}[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < -1.645 - \frac{5-6}{0.4}] \\
&= \mathbb{P}[Z < 0.855] \quad (\textbf{CLT}) \\
&= 0.8023
\end{aligned}$$

So the test has about an 80% chance of correctly rejecting the null hypothesis.

Often one picks α and β, and then those determine n. For example, to obtain NIH funding to run a clinical trial, you might decide to use $\alpha = 0.01$ and you want power $1 - \beta = 0.9$ for detecting an increase in average lifespan of 1 year.

You know that the average U.S. life expectancy is 77.6 years, with a standard deviation of about 14.5 years. (Is this well-posed?)

You want to show that your drug extends lives. The hypotheses are:

$$H_o : \mu_D \leq 77.6 \qquad \textbf{vs.} \qquad H_A : \mu_D > 77.6.$$

The test statistic is

$$ts = \frac{\bar{X} - 77.6}{14.5/\sqrt{n}}.$$

and the critical value is $z_{0.99} = 2.33$.

$$\begin{aligned}\textbf{power} = 0.9 &= 1 - \mathbb{P}[ts < 2.33] \\ &= \mathbb{P}[\frac{\bar{X} - 77.6}{14.5/\sqrt{n}} > 2.33] \\ &= \mathbb{P}[\frac{\bar{X} - 78.6 + 78.6 - 77.6}{14.5/\sqrt{n}} > 2.33] \\ &= \mathbb{P}[\frac{\bar{X} - 78.6}{14.5/\sqrt{n}} + \frac{1}{14.5/\sqrt{n}} > 2.33] \\ &= \mathbb{P}[Z > 2.33 - \frac{1}{14.5/\sqrt{n}}].\end{aligned}$$

From the z-table, $0.9 = \mathbb{P}[Z > -1.28]$, so

$$-1.28 = 2.33 - \frac{1}{14.5/\sqrt{n}}.$$

Solving shows that the least **integer** that achieves this power is $n = 2740$.

Some meta-points:

- Hypothesis testing is much like setting a confidence interval. A two-sided test of $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$ for a given α is often equivalent to whether or not a two-sided $(1 - \alpha)100\%$ confidence interval contains θ_0 (and similarly for one-sided tests and one-sided intervals).
- With large samples, one can get a statistically significant result that is of no practical importance.
- You must pick your null and alternative hypotheses before seeing the data. Also, you must pick two of α, β and n before looking at the data. Doing otherwise is cheating.
- If you make 100 tests of two identical groups, all with level $\alpha = 0.05$, you expect about 5 falsely significant results. See `http://xkcd.com/882/`.

15.3 Bayesian Hypothesis Testing

The **frequentist paradigm** treats unknown parameters as constants. To test hypotheses about parameters, a frequentist specifies a null and alternative hypothesis, draws a sample, and finds the probability of obtaining so extreme a sample when the null is true.

The **Bayesian paradigm** treats unknown parameters as random variables. To test hypotheses about parameters, a Bayesian has a prior belief about the unknown parameter, and specifies the null and alternative hypothesis. Then the Bayesian draws a sample and calculates the posterior probability of the null given the sample.

Recall Bayes' Theorem:

$$P(A_1|B) = \frac{P(B|A_1) * P(A_1)}{\sum_{i=1}^{k} P(B|A_i) * P(A_i)}$$

where the $A_1, \ldots, A_k$ are mutually exclusive and

$$P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_k) = 1.$$

This is a formalism for how we learn. $P(A_1)$ the **prior probability** of A_1, before observing B. Then we combine our prior probability with the new information on B, through $P(B|A_1)$, to get our new opinion, or the **posterior probability** of A_1, written as $P(A_1|B)$.

In the following example, we let A_i be the event that a certain probability is $i/10$, for $i = 1, \ldots, 9$.

15.4 RU486 Example

The "morning after" contraceptive RU486 was tested in a clinical trial in Scotland. This discussion slightly simplifies the design.

Assume 800 women report to a clinic; they have each had sex within the last 72 hours. Half are randomly assigned to take RU486; half are randomly given the conventional theory (high doses of estrogen and synthetic progesterone).

Among the RU486 group, none became pregnant. Among the conventional therapy group, there were 4 pregnancies. Does this show that RU486 is more effective than conventional treatment?

We shall compare the frequentist and Bayesian approaches.

Frequentist: Let p be the probability that an observed pregnancy came from an RU486 mother. If the two therapies are equally effective, then this is 0.5. A frequentist would test

$$\mathbf{H}_0 : p \geq 0.5 \quad \textbf{vs.} \quad \mathbf{H}_A : p < 0.5$$

If the P-value is small, then RU486 is deemed more effective than conventional treatment.

One has $n = 4$ observations from a binomial with probability p. One could use the Chinese menu, part II.c, but the sample size is small and so we can calculate the binomial probabilities exactly:

$$P - \textbf{value} = P[\ \textbf{0 successes in 4 tries}\ |\mathbf{H}_0\ \textbf{true}\ ] = (1 - .5)^4 = 0.0625.$$

Most frequentists would fail to reject, since $.0625 > .05$.

Bayesian: A Bayesian begins with a prior over all the possible values for p. For example, suppose we thought we had no information a priori about the probability that a child came from the RU486 group. In that case all values of p between 0 and 1 would be equally likely and our prior on p is the uniform distribution on [0,1], or Beta(1,1).

But the idea may be more clear without using the Bayes-binomial trick. So we approximate Beta(1,1) by assuming that each of the following values for p is equally likely: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. So each value has prior probability 1/9.

If we picked one of the values, say $p = .1$, then that means the probability of a pregnancy coming from the RU486 group is 0.1, so 0.9 is the chance it comes from the conventional group. But we do not know which value is correct.

Therefore we use Bayes theorem to find the posterior probability of each of the 9 possible values.

Value	**Prior**	**P(data | value)**	**Product**	**Posterior**
p	**P[value]**	**P[**$k = 0 \mid p$**]**		**P[value | data]**
.1	1/9	.656	.0729	.427
.2	1/9	.410	.0455	.267
.3	1/9	.240	.0266	.156
.4	1/9	.130	.0144	.084
.5	1/9	.063	.0070	.041
.6	1/9	.026	.0029	.017
.7	1/9	.008	.0009	.005
.8	1/9	.002	.0002	.001
.9	1/9	.000	.0000	.000
	1		**0.1704**	1

The most likely value is $p = 0.1$, with posterior probablity 0.427.

And the posterior probability of the alternative hypothesis, $\mathbf{H}_A : p < 0.5$, is $0.427 + 0.267 + 0.156 + 0.084 = 0.934$.

Note that in performing the Bayes calculation,

- We were able to find the probability that $p < 0.5$, which we could not do in the frequentist framework.
- In calculating this, we used only the data that were observed. Data that were more extreme than what we observed plays no role in the calculation or the logic.

Suppose a different Bayesian analyzes the same data. But their prior does not put equal weight on the 9 models; they put prior weight 0.52 on the model $p = .5$ and equal weight on the others.

Value	Prior	P(data | value)	Product	Posterior
p	**P**[value]	**P**[$k = 0 \mid p$]		**P**[value | data]
.1	.06	.656	.0394	.326
.2	.06	.410	.0246	.204
.3	.06	.240	.0144	.119
.4	.06	.130	.0078	.064
.5	.52	.063	.0325	.269
.6	.06	.026	.0015	.013
.7	.06	.008	.0005	.004
.8	.06	.002	.0001	.001
.9	.06	.000	.0000	.000
	1		**0.1208**	1

So this Bayesian has posterior probability of the alternative as 0.713.