Statistical Hypothesis Testing

Schwartz

August 29, 2016

A Brief History of Statistics (Part I)

Significance testing is largely the product of Karl Pearson (1857–1936), William Sealy Gosset (1876–1937), and Ronald Fisher (1890–1962), although evidence of its use dates back to Laplace (1749–1827) in the 1770's. Pearson created the notion of a p-value and (Pearson's) chi-squared test and founded the world's first statistics department at University College London in 1911. Gosset developed and penned the t-distribution and t-test under the pseudonym Student due to the objections of his employer – the original Guinness Brewery in Dublin, Ireland – regarding publication of internal practices. And Fisher created analysis of variance and popularized the notions of null hypothesis and significance test. In addition to being regarded as the father of modern statistical science and experimental design, Fisher also made significant contributions to agricultural biology and genetics. Indeed, Richard Dawkins named him "the greatest biologist since Darwin".

Hypothesis testing was developed by Jerzy Neyman (1894 – 1981) and Egon Pearson (1895–1980, son of Karl Pearson). Building on these ideas, Neyman later introduced confidence intervals into the statistics landscape. At the time of the publication of their work on hypothesis testing in 1933, Neyman and Pearson (along with Fisher) were faculty members at the University College London in the department of statistics (founded by the older Pearson). While Fisher as a result of his agricultural background emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions, Neyman (who teamed with the younger Pearson) emphasized mathematical rigor and methods to obtain more results from many samples and a wider range of distributions.

Initially a Bayesian, Fisher but sought to provide a more "objective" approach to inference. The significance testing he developed did not use the notion of an alternative hypothesis – only a null hypothesis – and hence did not involve the notion of Type II error. Fisher's interpretation of p-values was informal: p-values were only meant to provide guidance for potential future experiments. Neyman and Pearson on the other hand formalized hypothesis testing with Type I/II errors and developed a procedure to choose between competing hypotheses. They considered their formulation to be an improved and more objective generalization of significance testing as it provided a decision making tool to determine researcher behavior without requiring any inductive inference on the part of the researcher.

A Brief History of Statistics (Part II)

Fisher and Neyman/Pearson clashed bitterly, and often. As they all shared the same building at the University College London they had ample opportunity to cross paths (and swords – although only Fisher was ever knighted – and not until many years later – and Neyman was, after all, Polish, not English). They disagreed about the proper role of models in statistical inference. Fisher thought the Neyman/Pearson approach was not applicable to scientific research because (1) initial assumptions about the null hypothesis are often discovered to be questionable as unexpected sources of error appear over the course of the experiment and (2) rigid reject/accept decisions based on models formulated before data is collected are incompatible with the real-world scenario faced by scientists and attempts to apply such formulations to scientific research would lead to mass confusion [as it has].

In 1938 Neyman left University College London and moved to the University of California, Berkeley. This put much of the planetary diameter between both his partnership with Pearson and his dispute with Fisher. A further respite in the debate was provided by World War II. Nonetheless, the disagreement between Fisher and Neyman only terminated (unresolved after 27 years) with Fisher's death in 1962. Neyman wrote a well-regarded eulogy of Fisher upon his death. And some of Neyman's later publications reported p-values and significance levels.

Afterword:

In an apparent effort to provide a "non-controversial" theory (as well as likely from confusion and misunderstanding of the topic, *per se*) the modern version of hypothesis testing used today is an inconsistent hybrid of the "Fisher versus Neyman/Pearson" formulations developed in the early 20th century. Rather than comparing two directly competing realistic hypotheses, one of the hypotheses is made to be a "no effect null hypothesis" so (despite great conceptual differences and caveats) p-values can be interpreted from both the Fisher and the Neyman/Pearson perspectives. Neyman and Pearson provided the stronger terminology, the more rigorous mathematics and the more consistent philosophy, but the hypothesis testing used today has more similarities with Fisher's method than theirs.

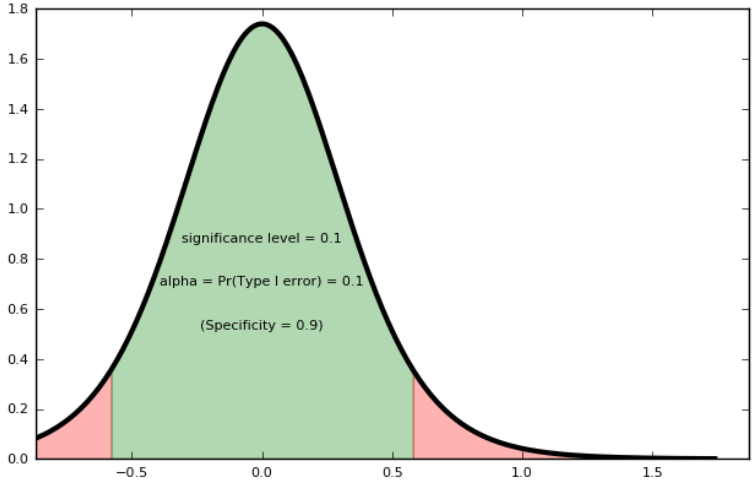
Outline

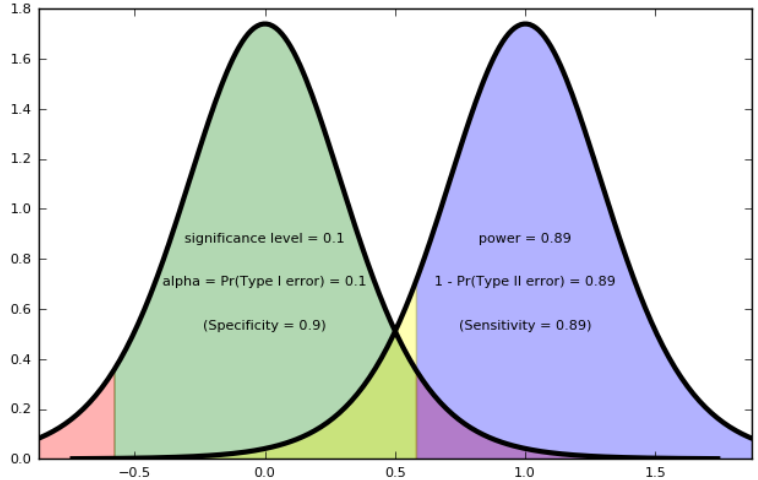
- Know what a **null hypothesis** is
  - Know what a **p-value** is

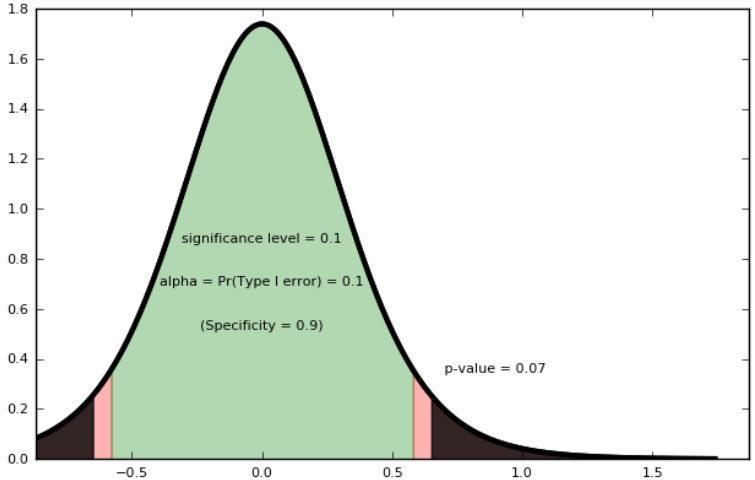
    (don't you *dare* mess this up EVER)
  - Know what an α-**significance level** is
    - Know what a **two-tailed versus one-tailed test** is
  - Know how this relates to **confidence intervals**
- Know what an **alternative hypothesis** is
  - Know what **power** is
- Know a sh*t-ton of statistical tests, like these ones:
  - Two sampled z/t-test, common variance
  - Two sampled z/t-test, unique variances
  - Paired samples t-test
  - Pearson's χ^2-test
  - Kolmogorov-Smirnov (K-S) test
  - F-test
- And be able to correctly apply them where appropriate
- And know yourself a little *Bonferroni* and *FDR*

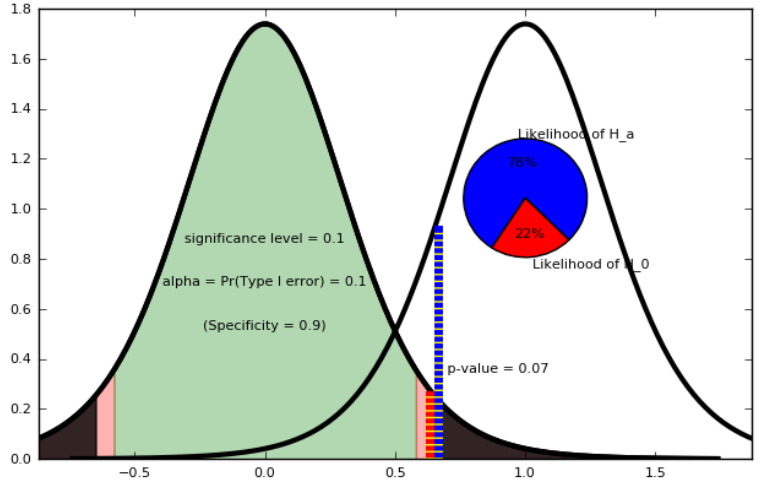
$$\Pr(\text{as or more extreme than what you saw} | H_0)$$

Hypothesis Testing

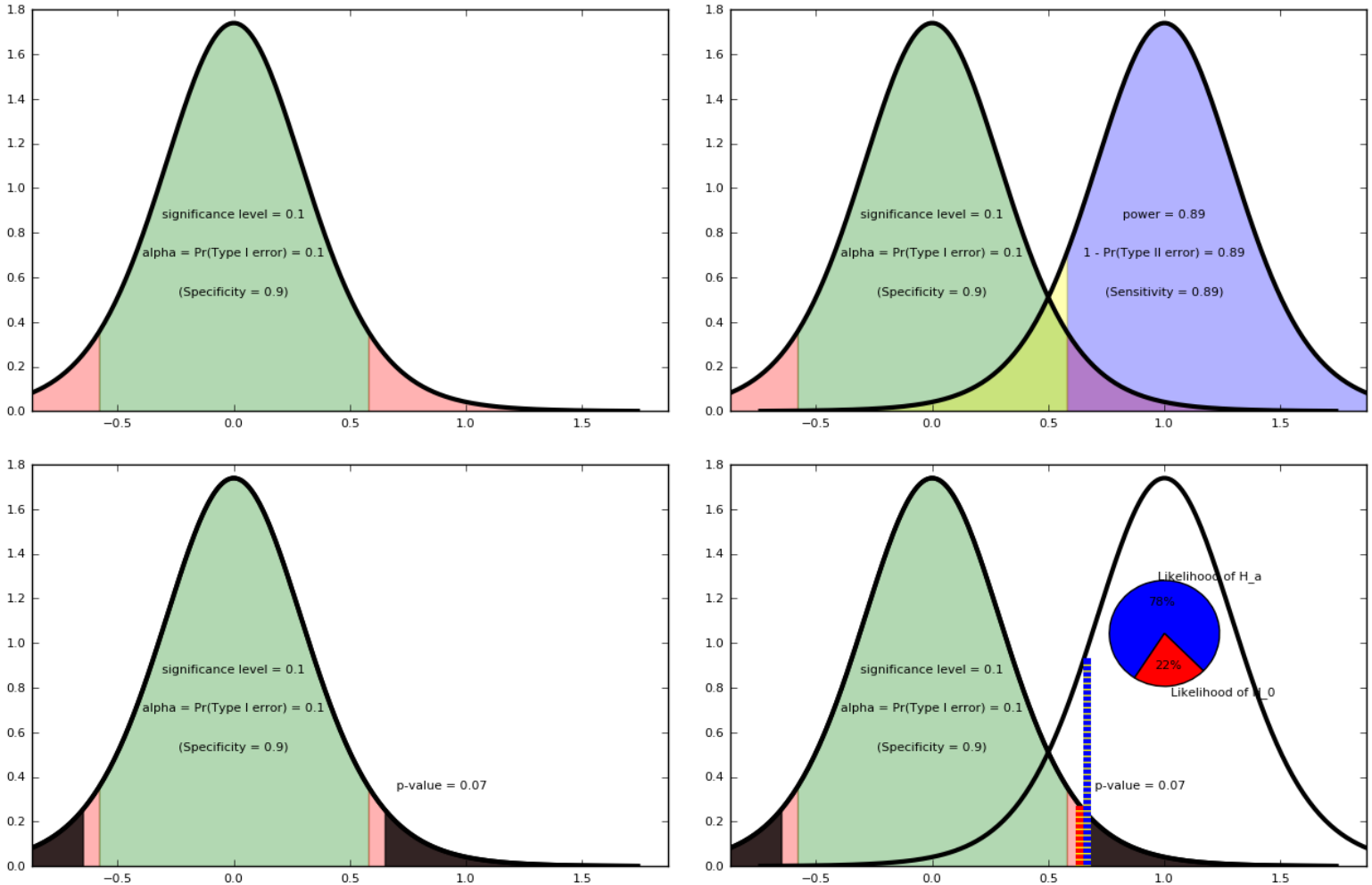






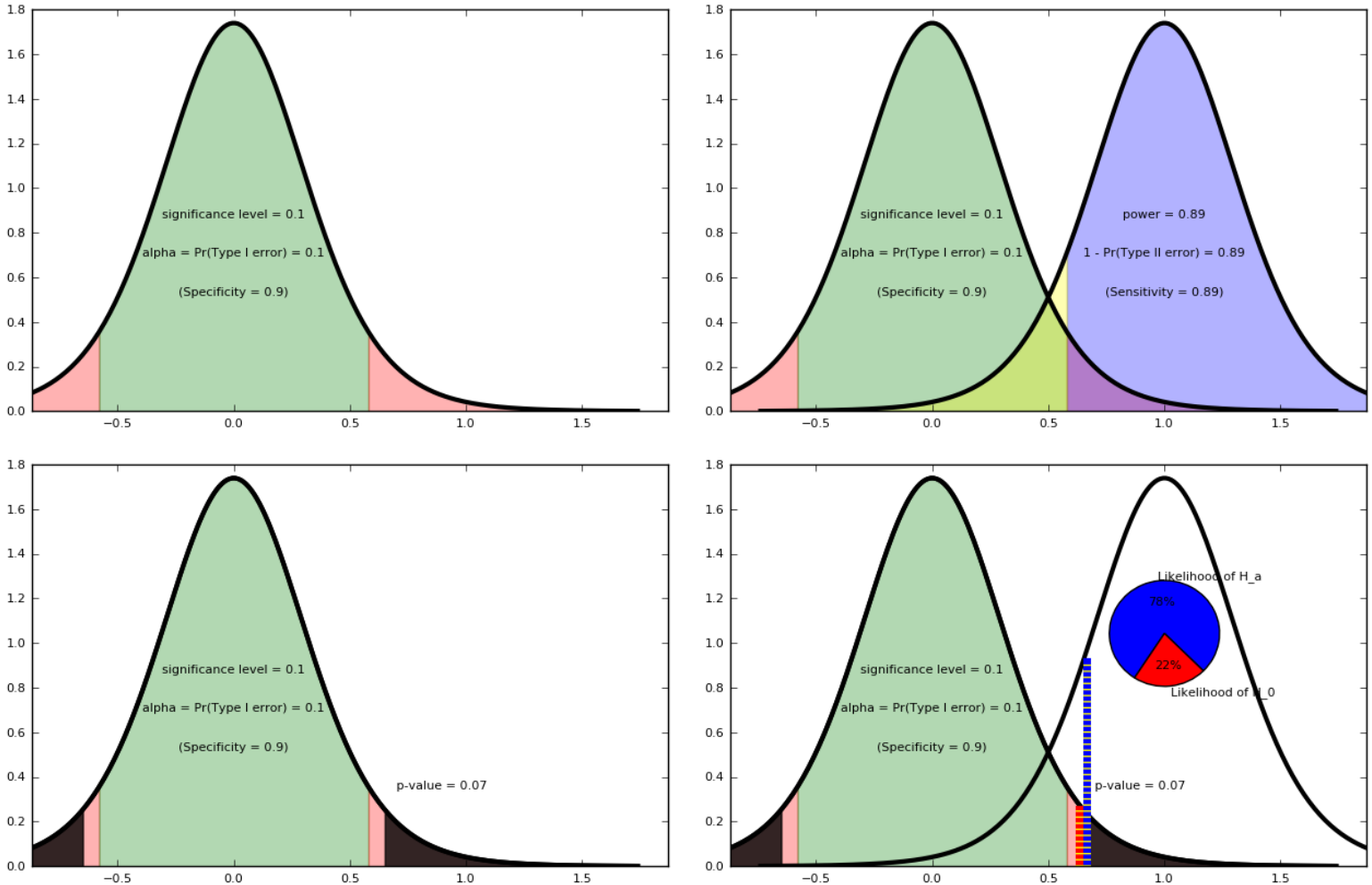


Hypothesis Testing



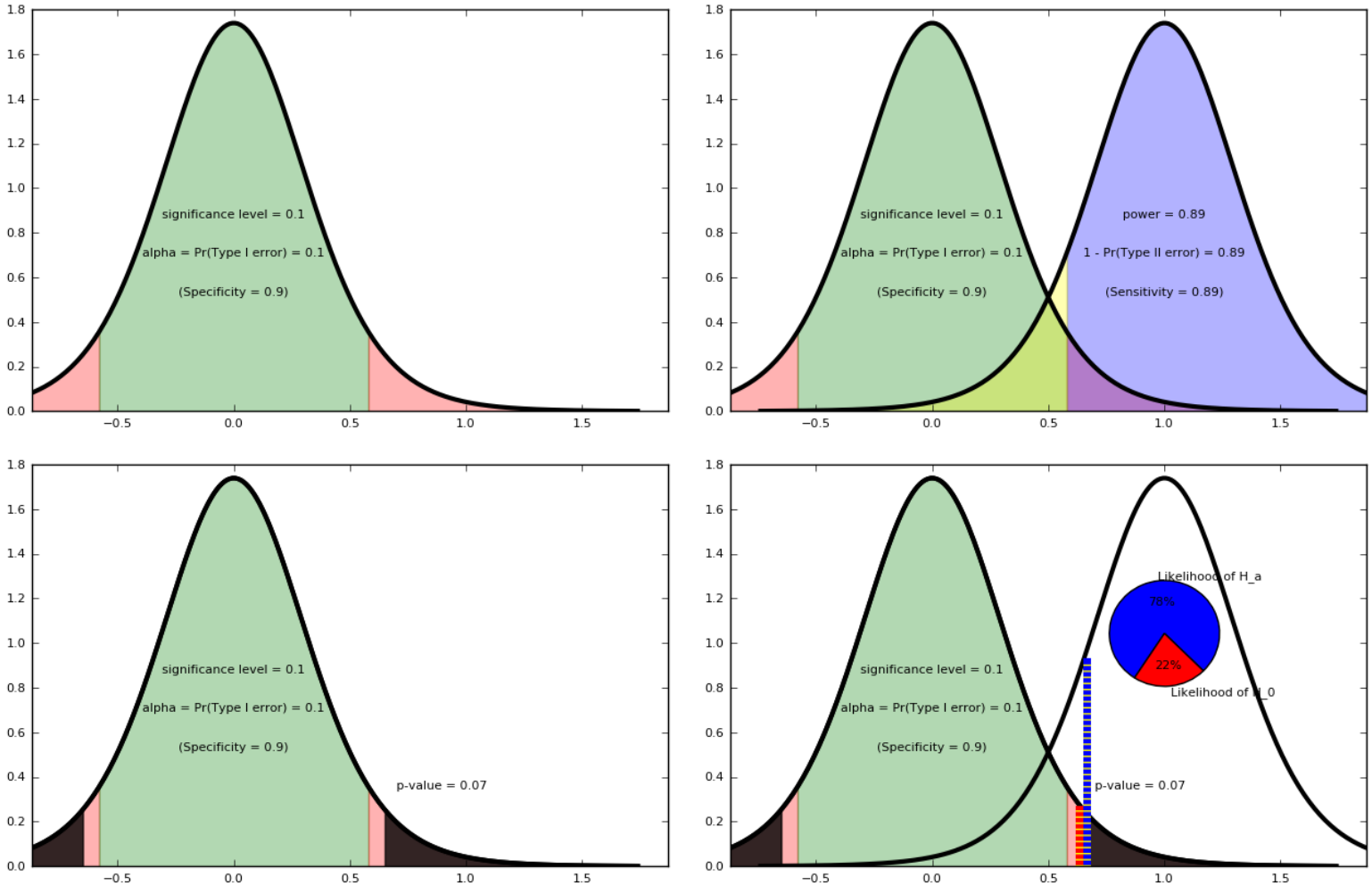
- What α significance level are we testing at?

Hypothesis Testing



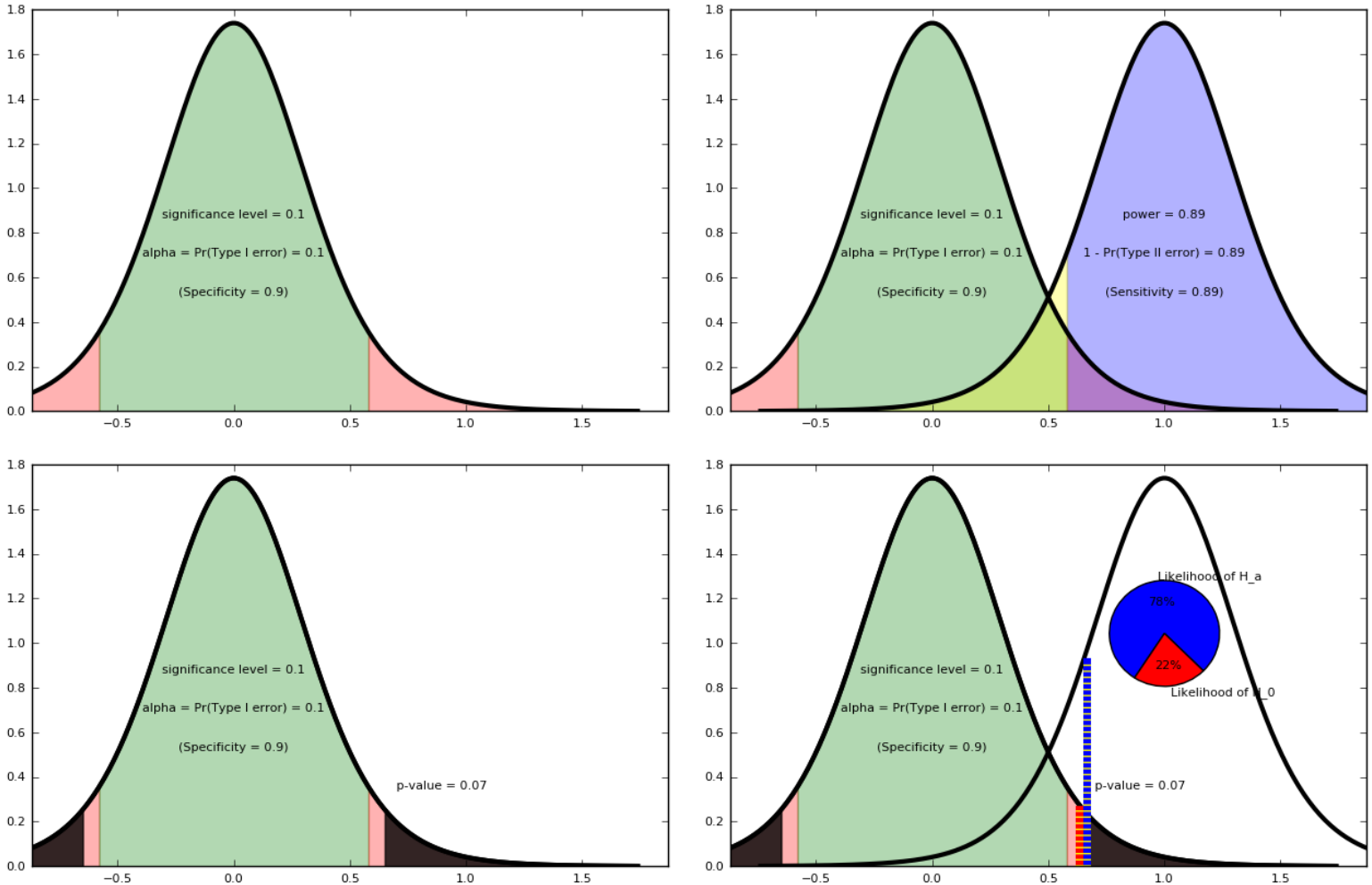
- How did we choose this α significance level?

Hypothesis Testing



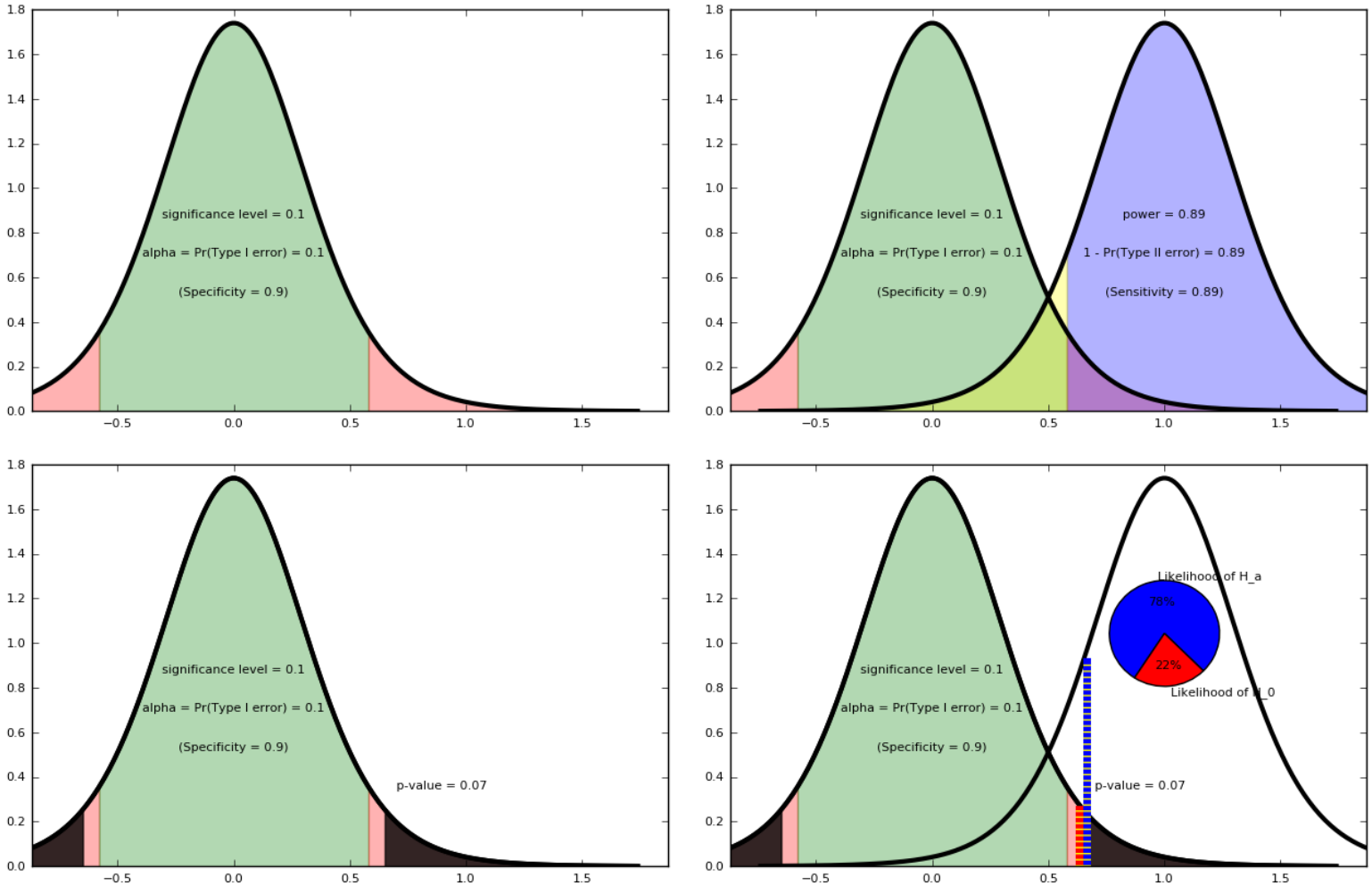
- How did we construct this α significance level?

Hypothesis Testing



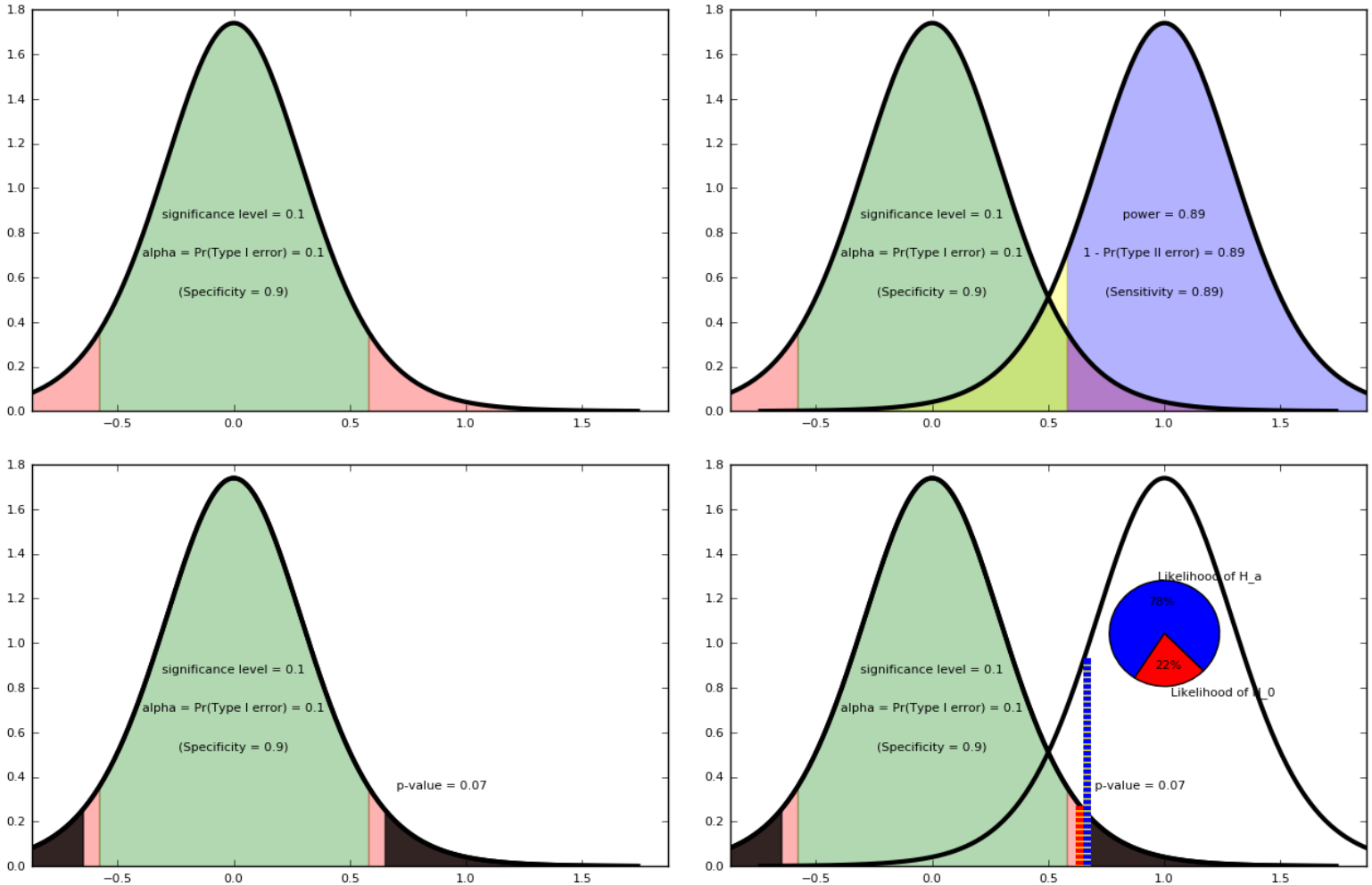
- We observed a p-value of 0.07... what's Pr(H_0 True)?

Hypothesis Testing



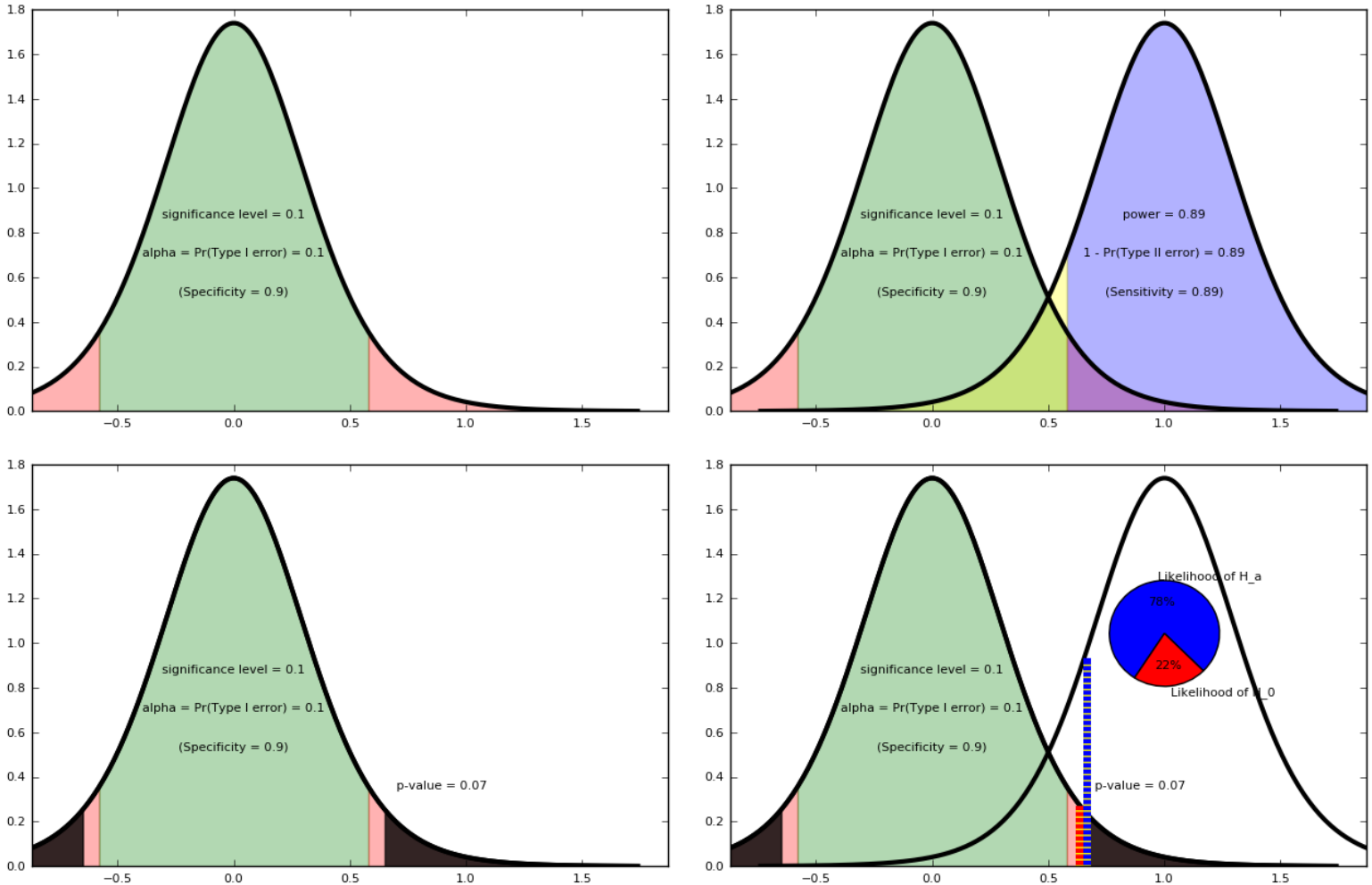
- We observed a p-value of 0.07... what's Pr(H_0 False)?

Hypothesis Testing



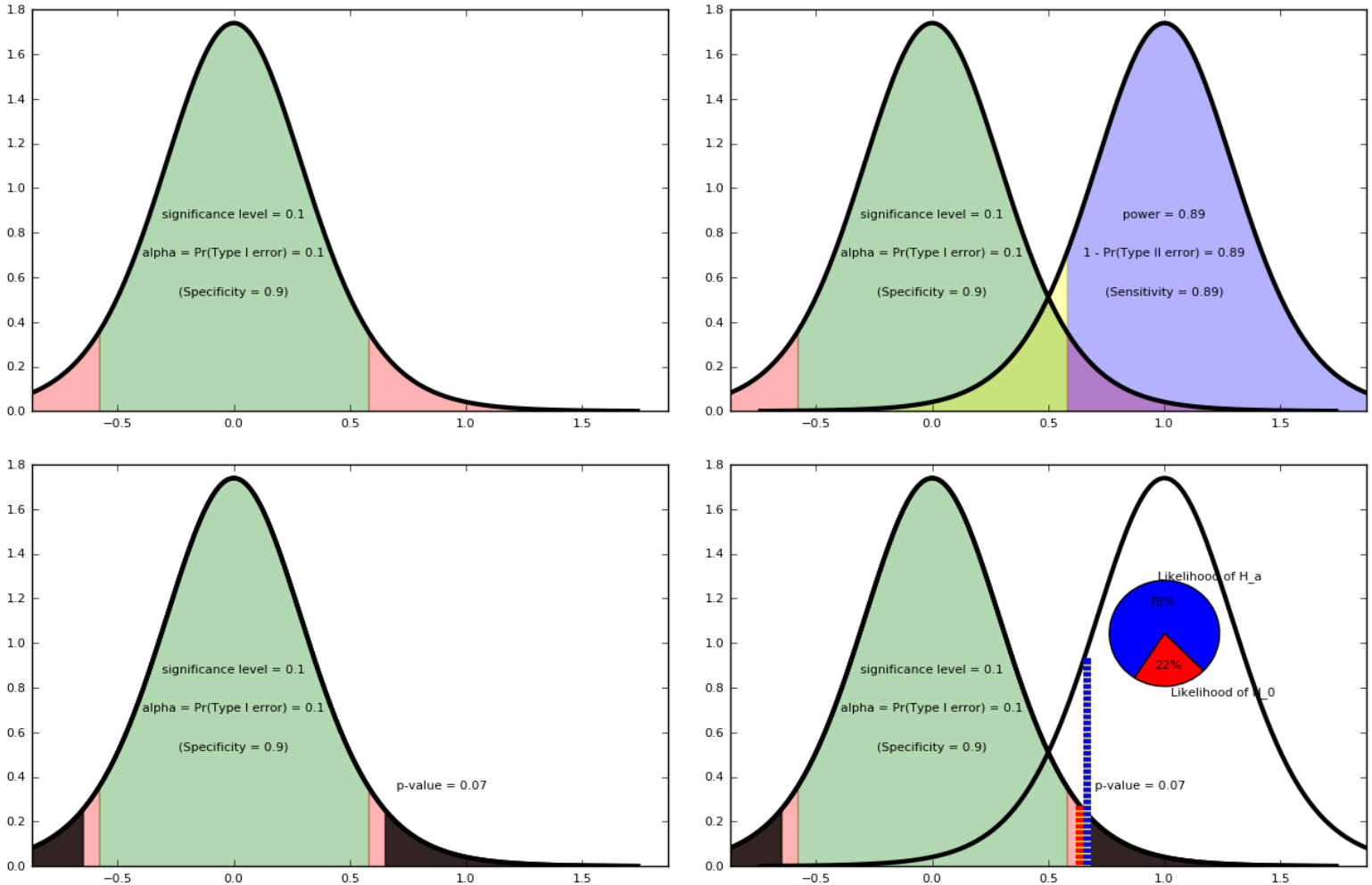
- We observed a p-value of 0.07... Pr(H_0 Incorrectly Rejected)?

Hypothesis Testing



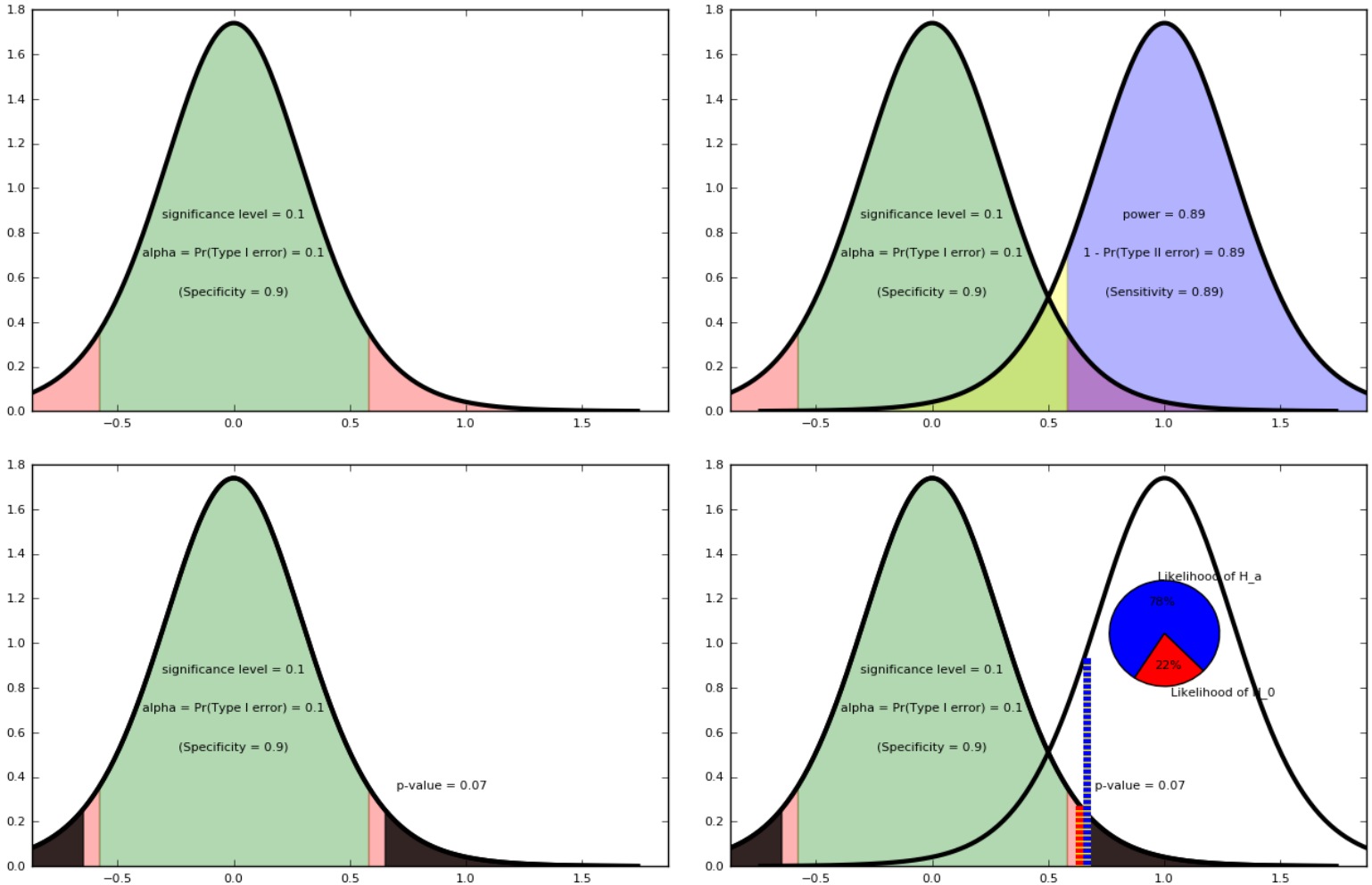
- We observed a p-value of 0.07... Pr(H_0 Correctly Accepted)?

Hypothesis Testing



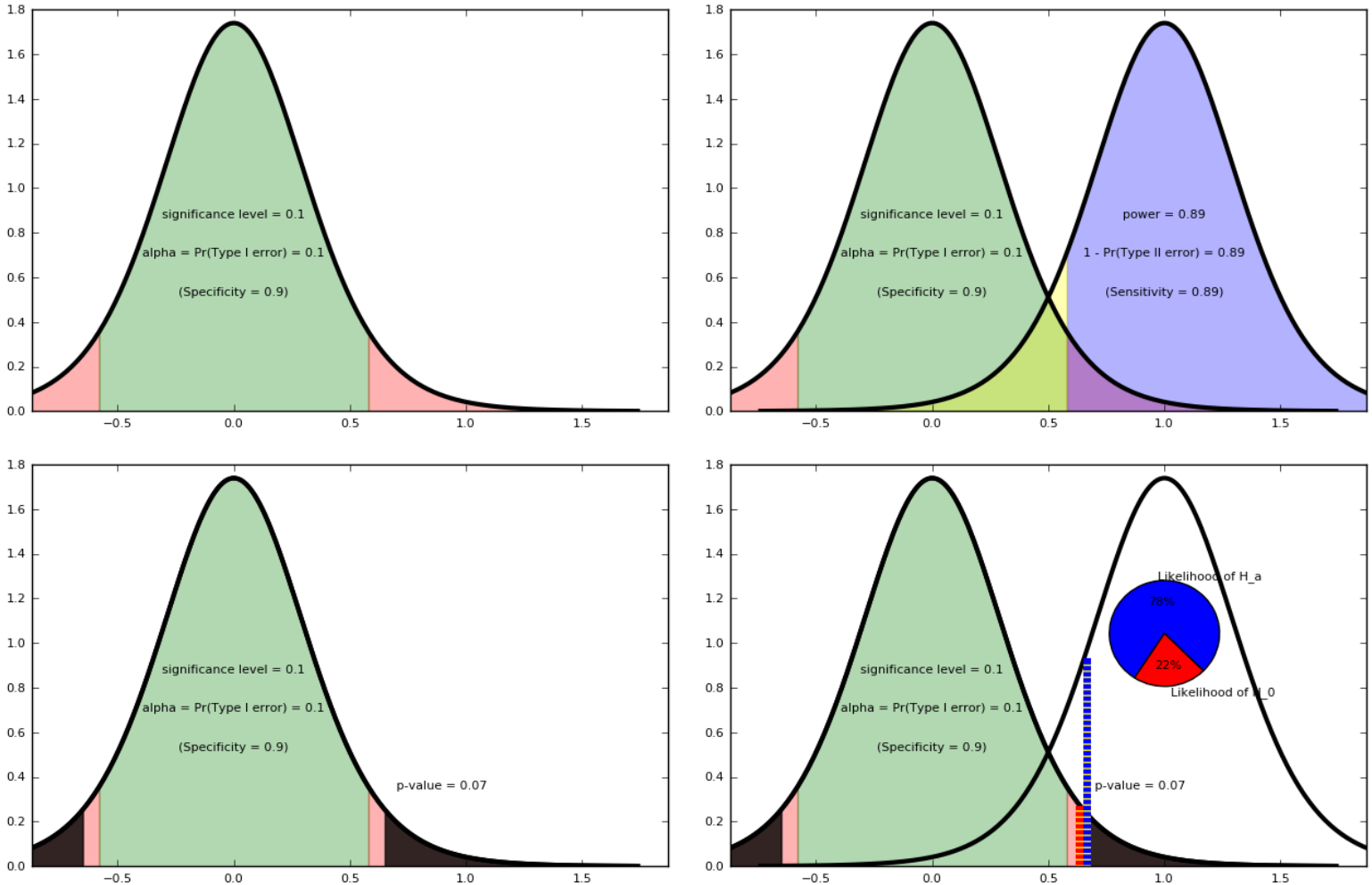
- $\alpha = 0.1$.... is there a 90% chance H_a is True?

Hypothesis Testing



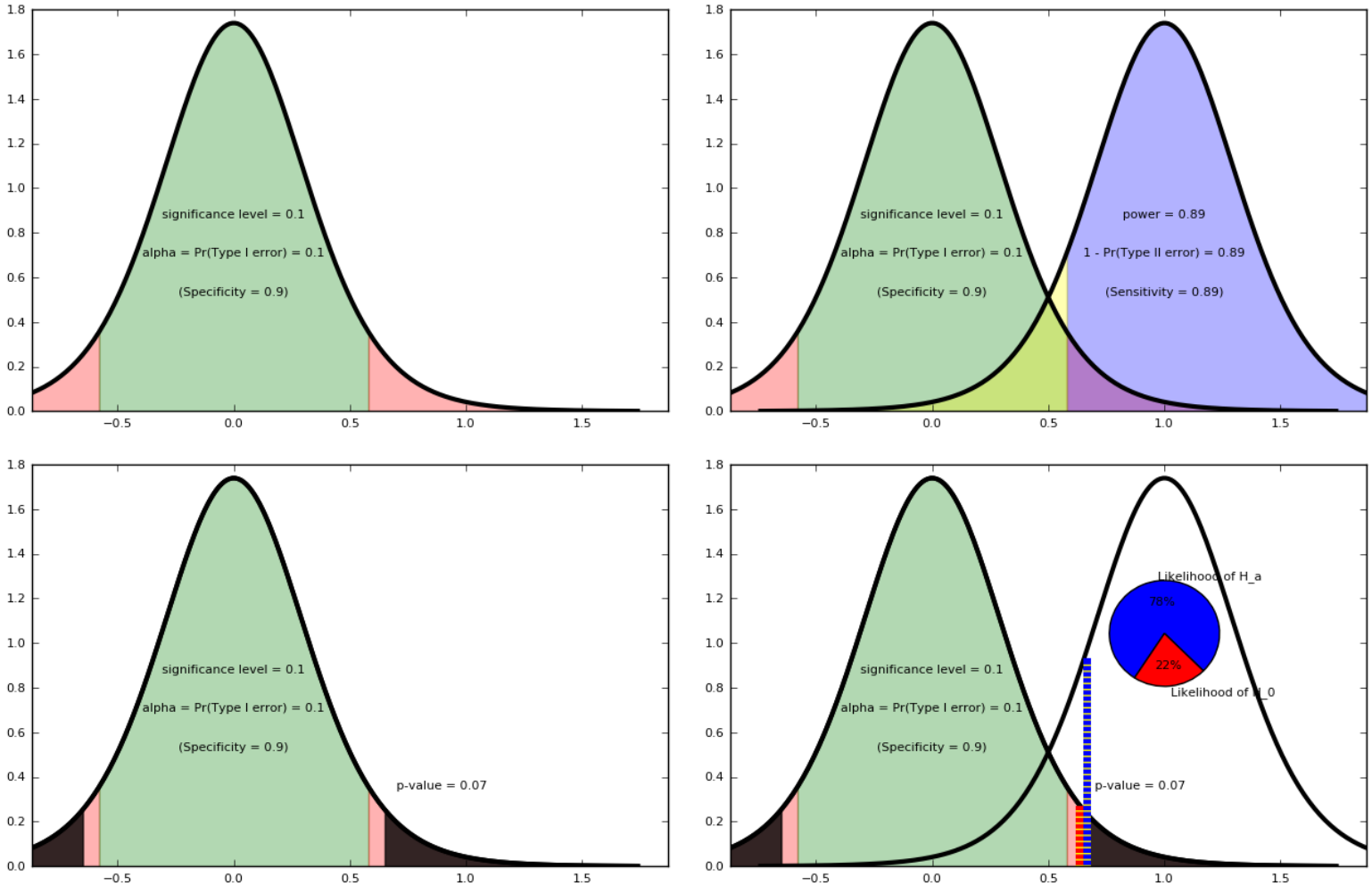
- $\alpha = 0.1$.... is there a 10% chance H_0 is False?

Hypothesis Testing



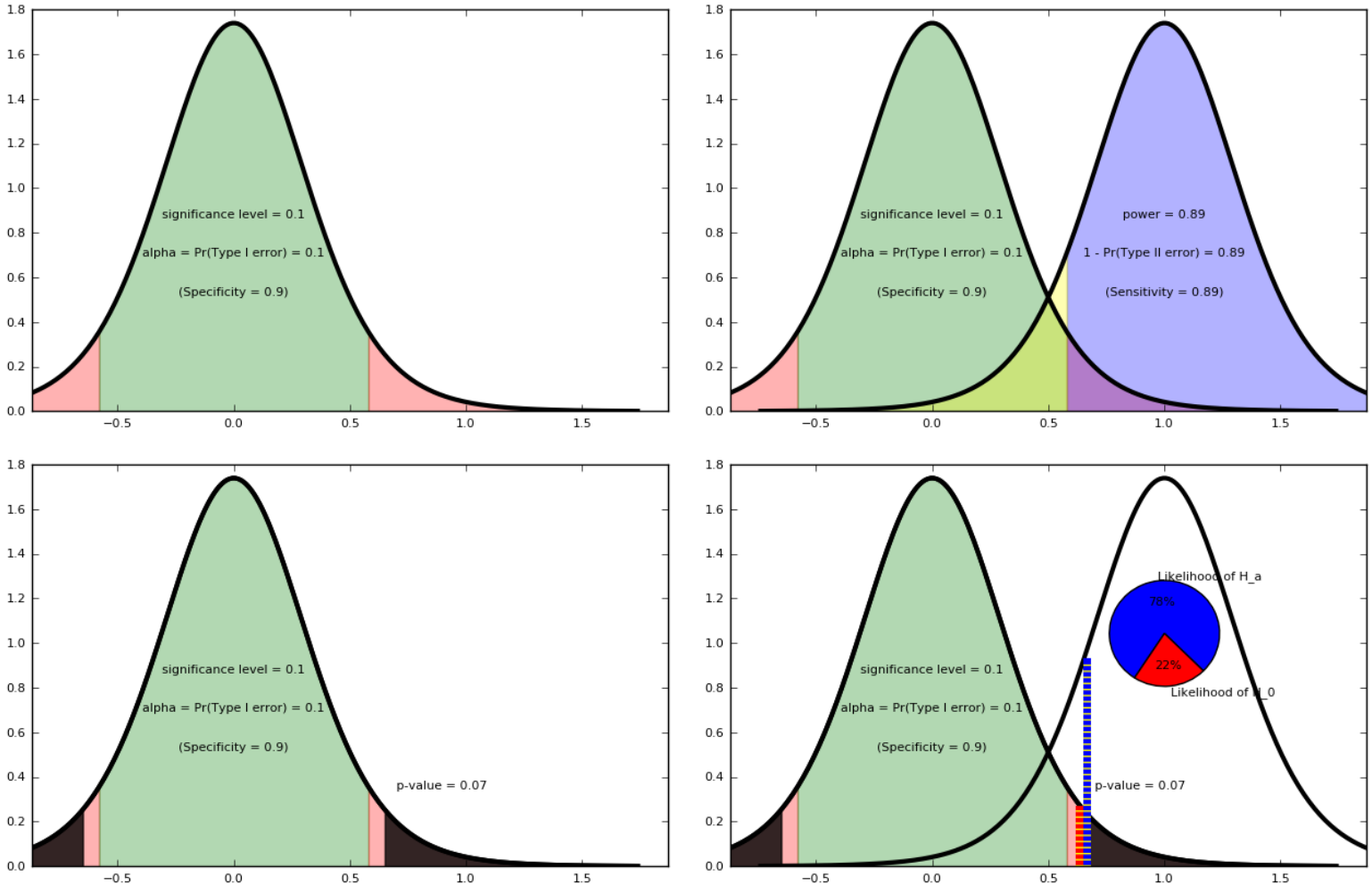
- $\alpha = 0.1$.... is there a 10% chance H_a is True?

Hypothesis Testing



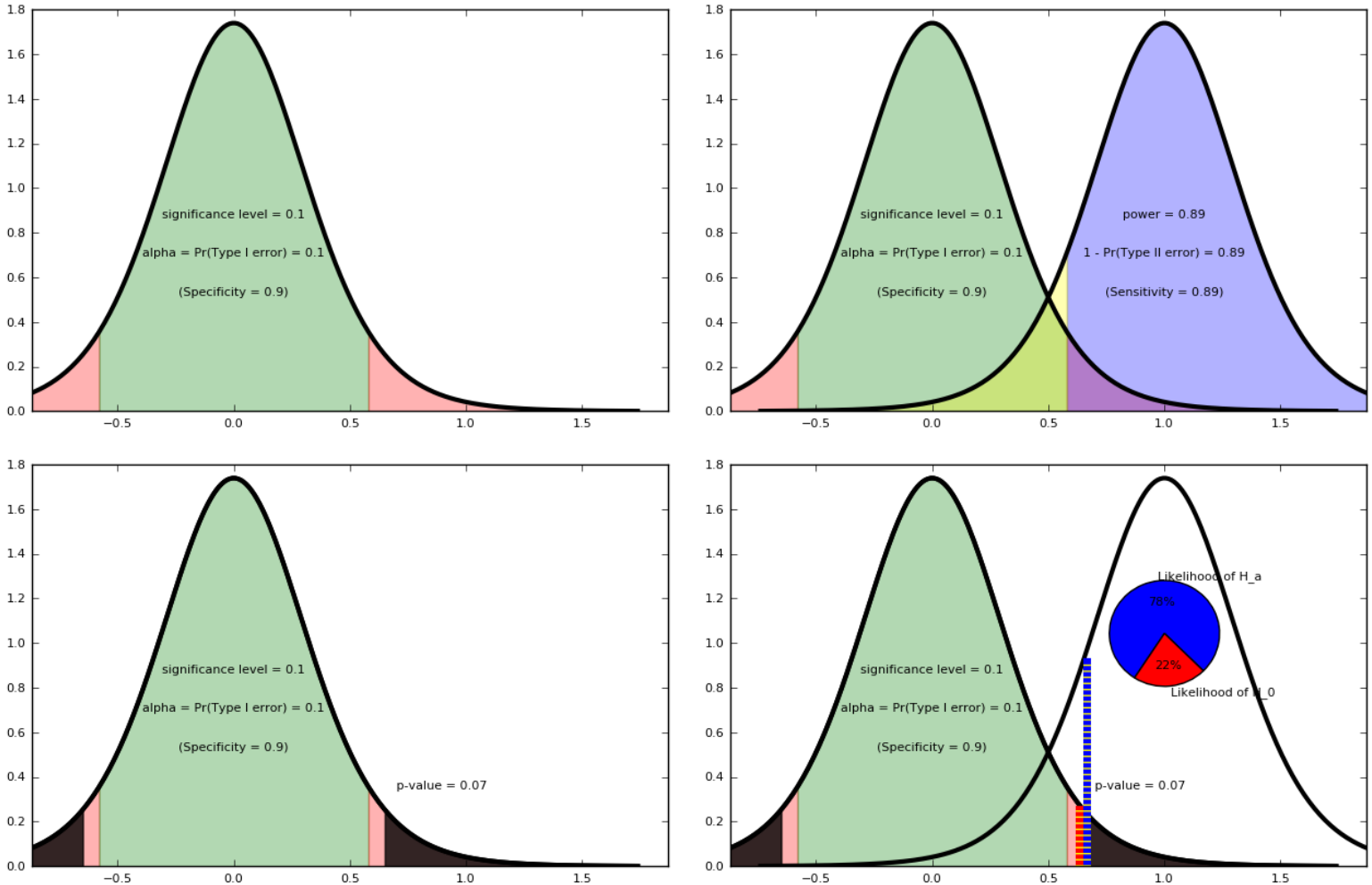
- $p = 0.07\ldots$ is there a 93% chance H_0 is True?

Hypothesis Testing



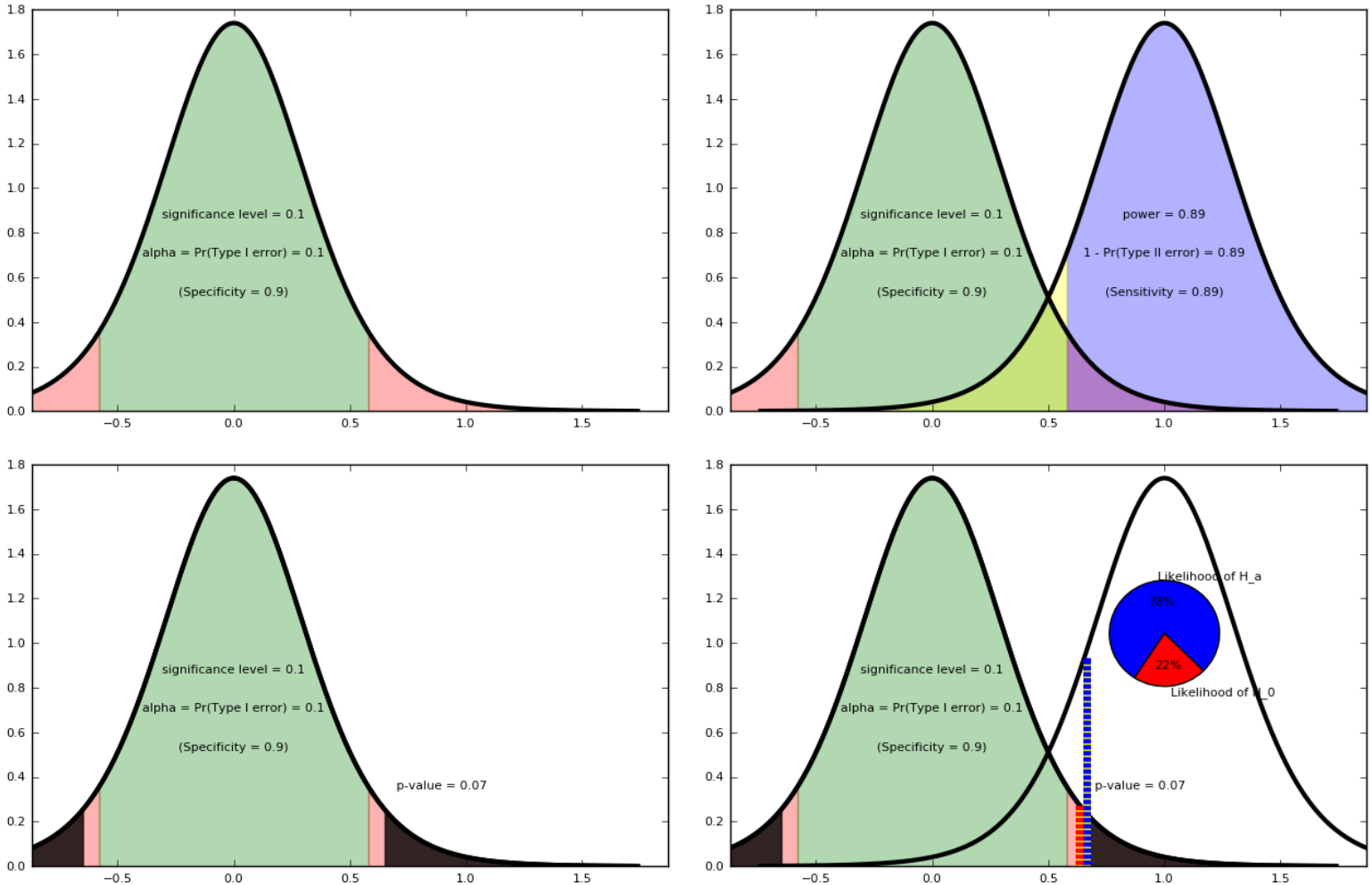
- $p = 0.07\ldots$ is there a 7% chance H_0 is False?

Hypothesis Testing



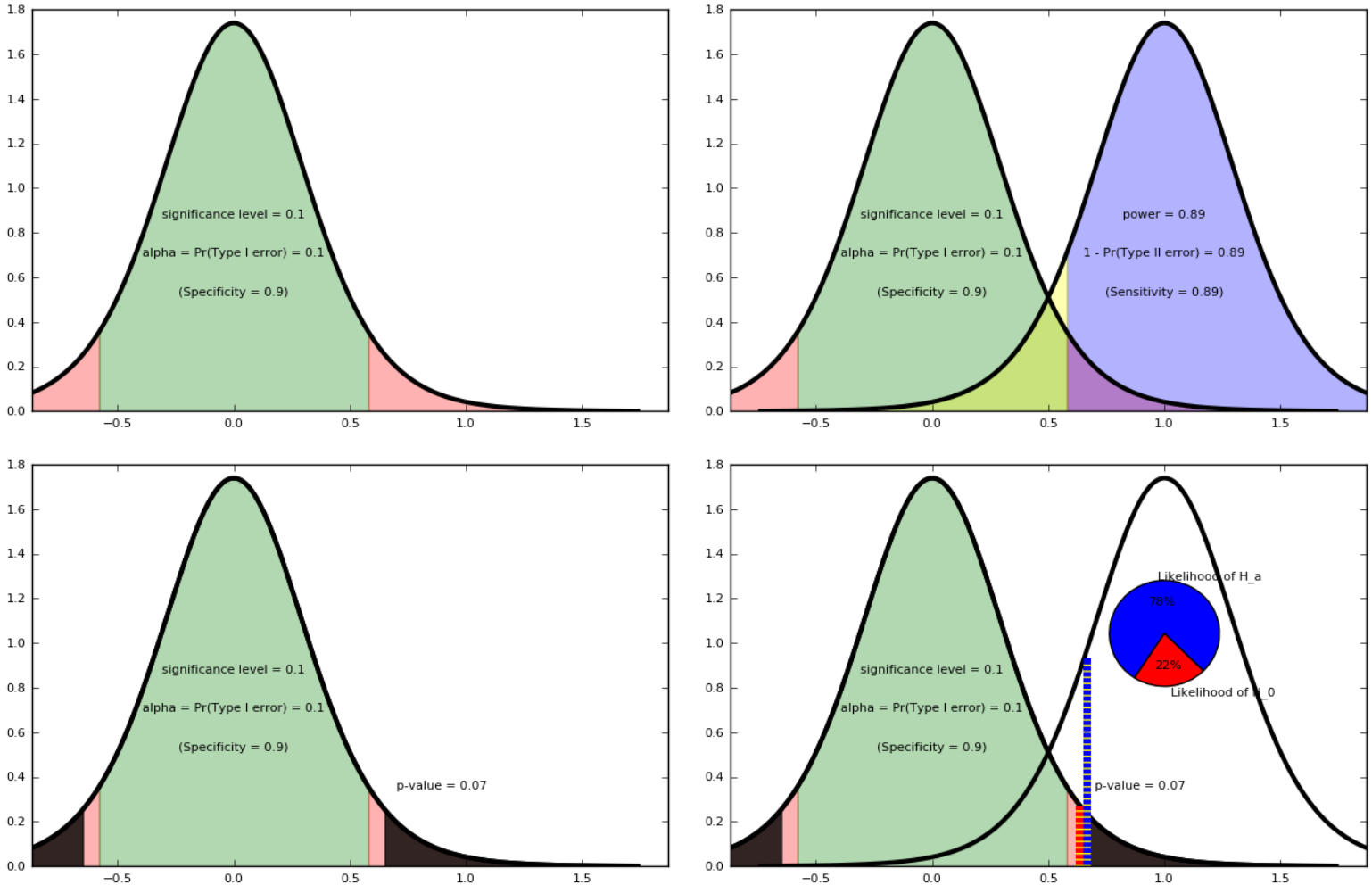
- $p = 0.07\ldots$ is there a 7% chance H_a is True?

Hypothesis Testing



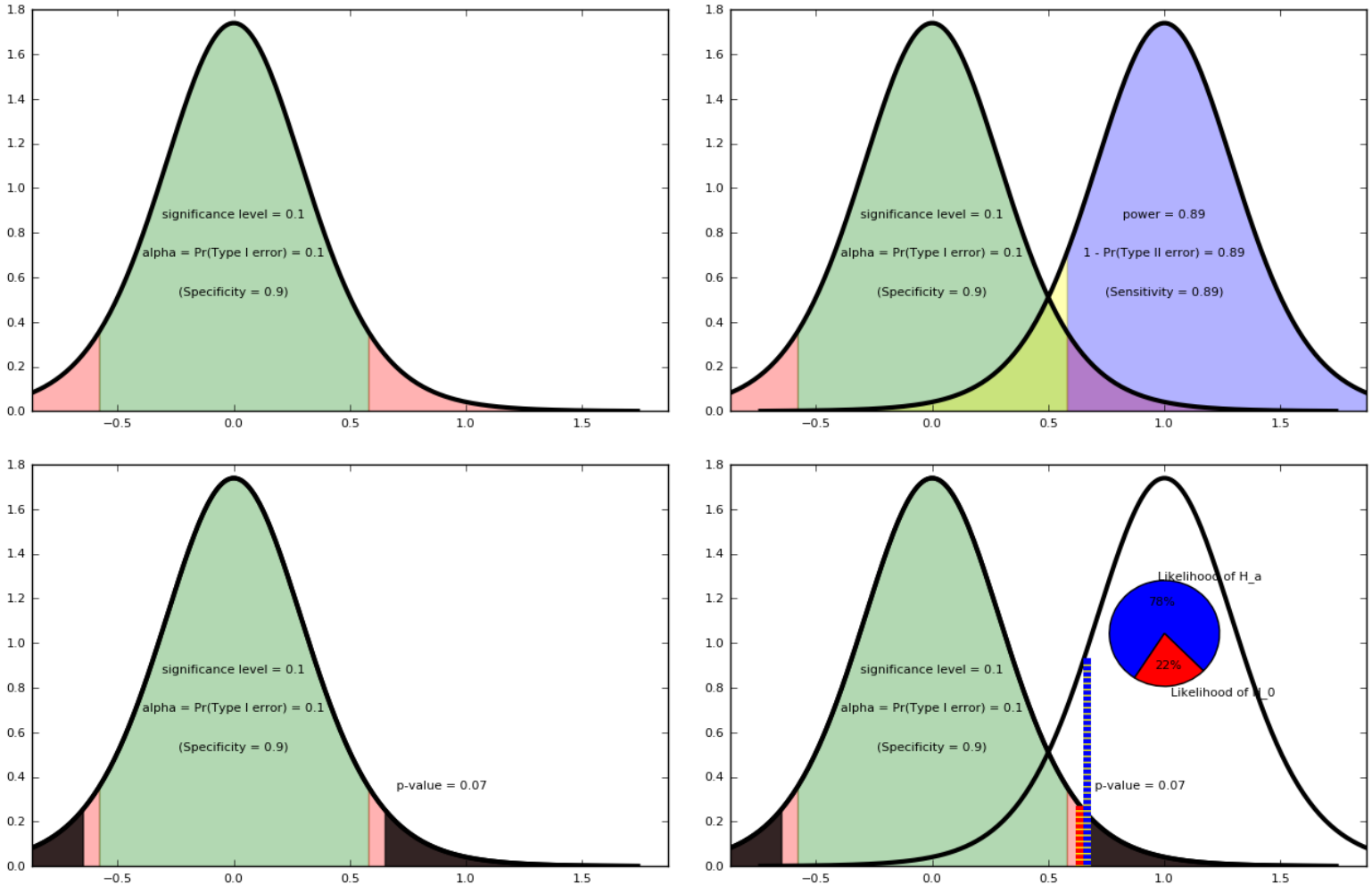
- $p = 0.07\ldots$ does $\Pr(H_0 \text{ Incorrectly Rejected}) = 0.07$?

Hypothesis Testing



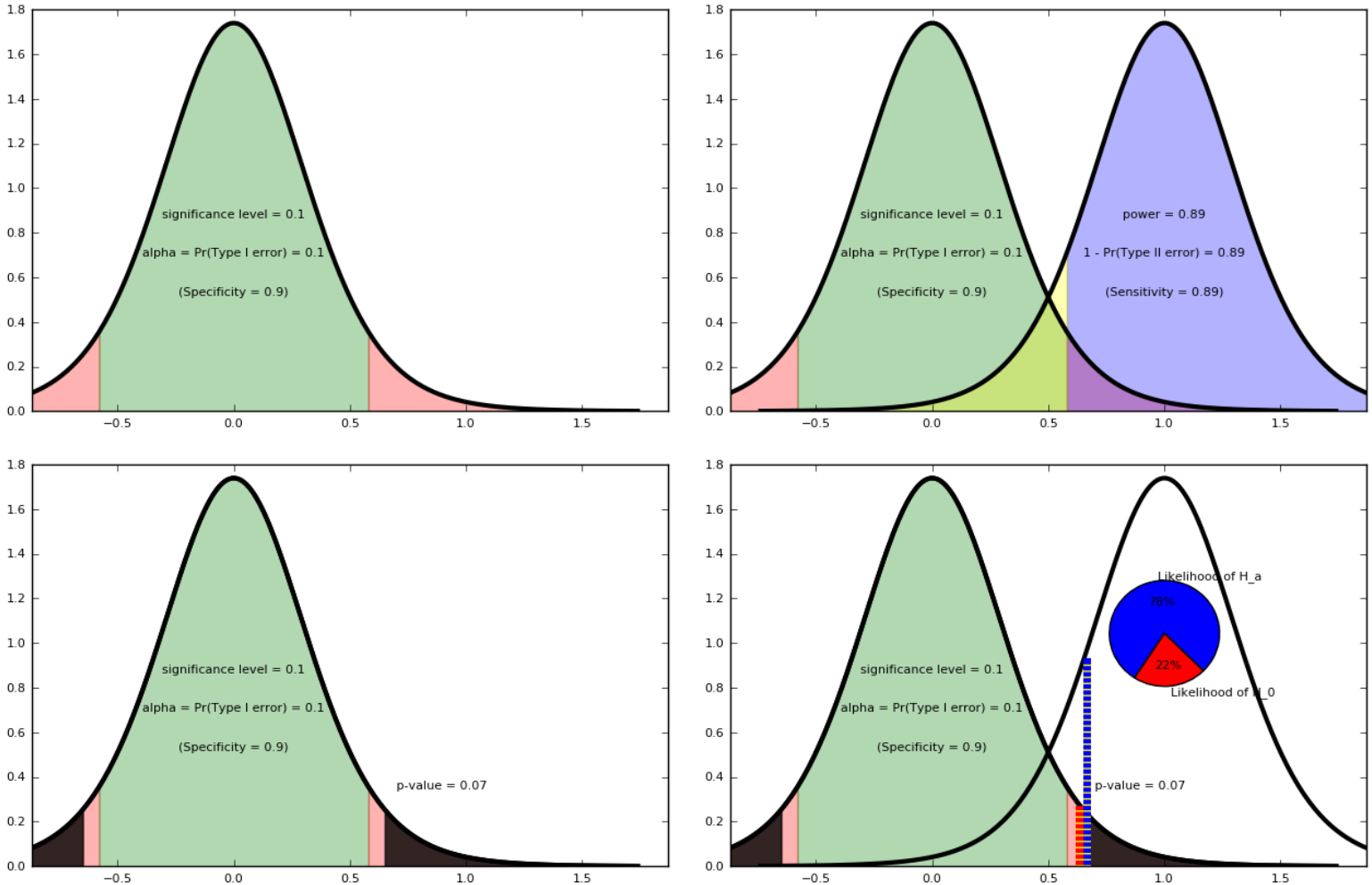
- $p = 0.07\ldots$ does $\Pr(H_0 \text{ Correctly Accepted}) = 0.93$?

Hypothesis Testing



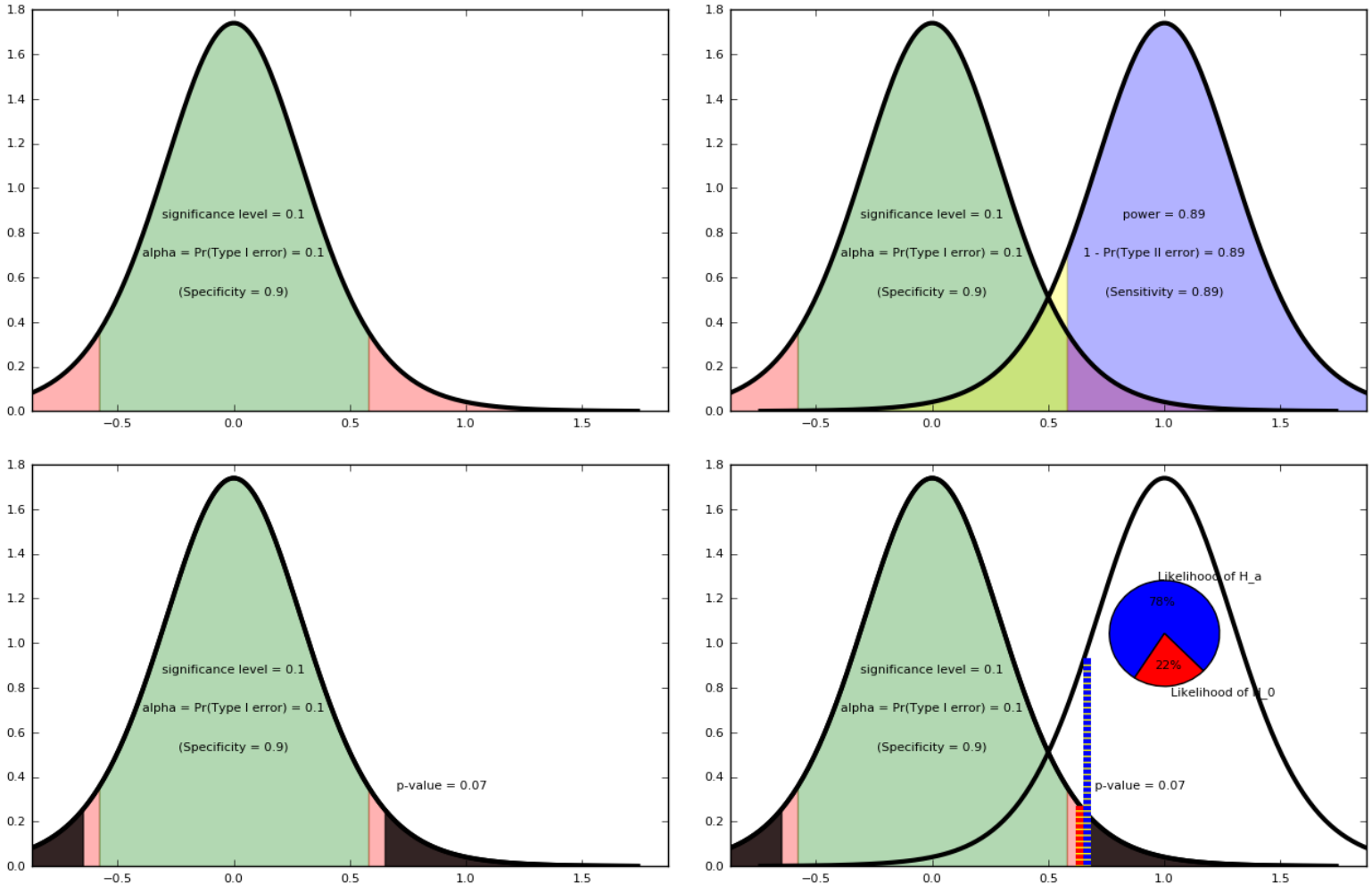
- What is Pr(H_0 Incorrectly Rejected)?

Hypothesis Testing



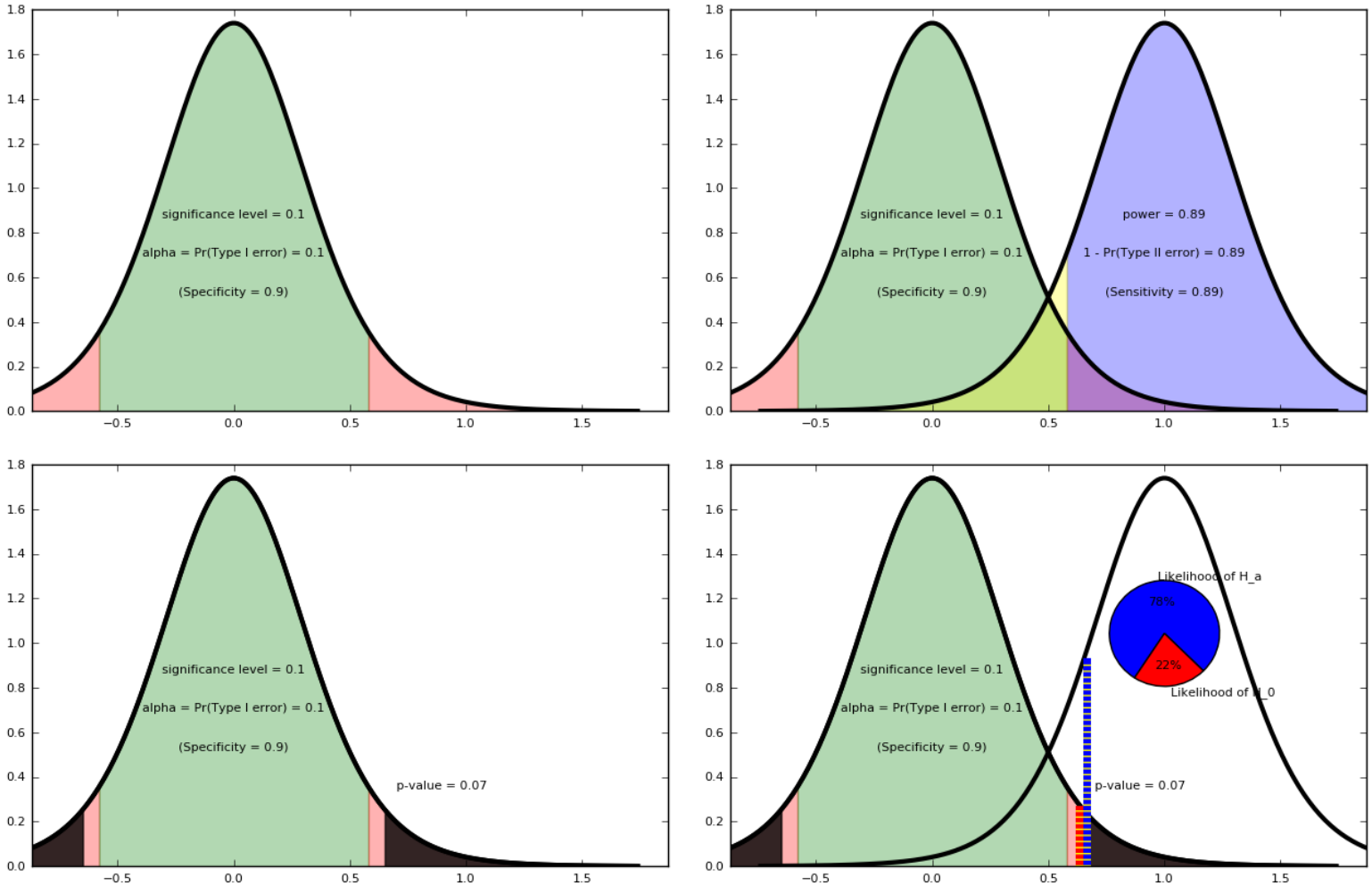
- What is $\Pr(H_0 \text{ Rejected} \mid H_0 \text{ True})$?

Hypothesis Testing



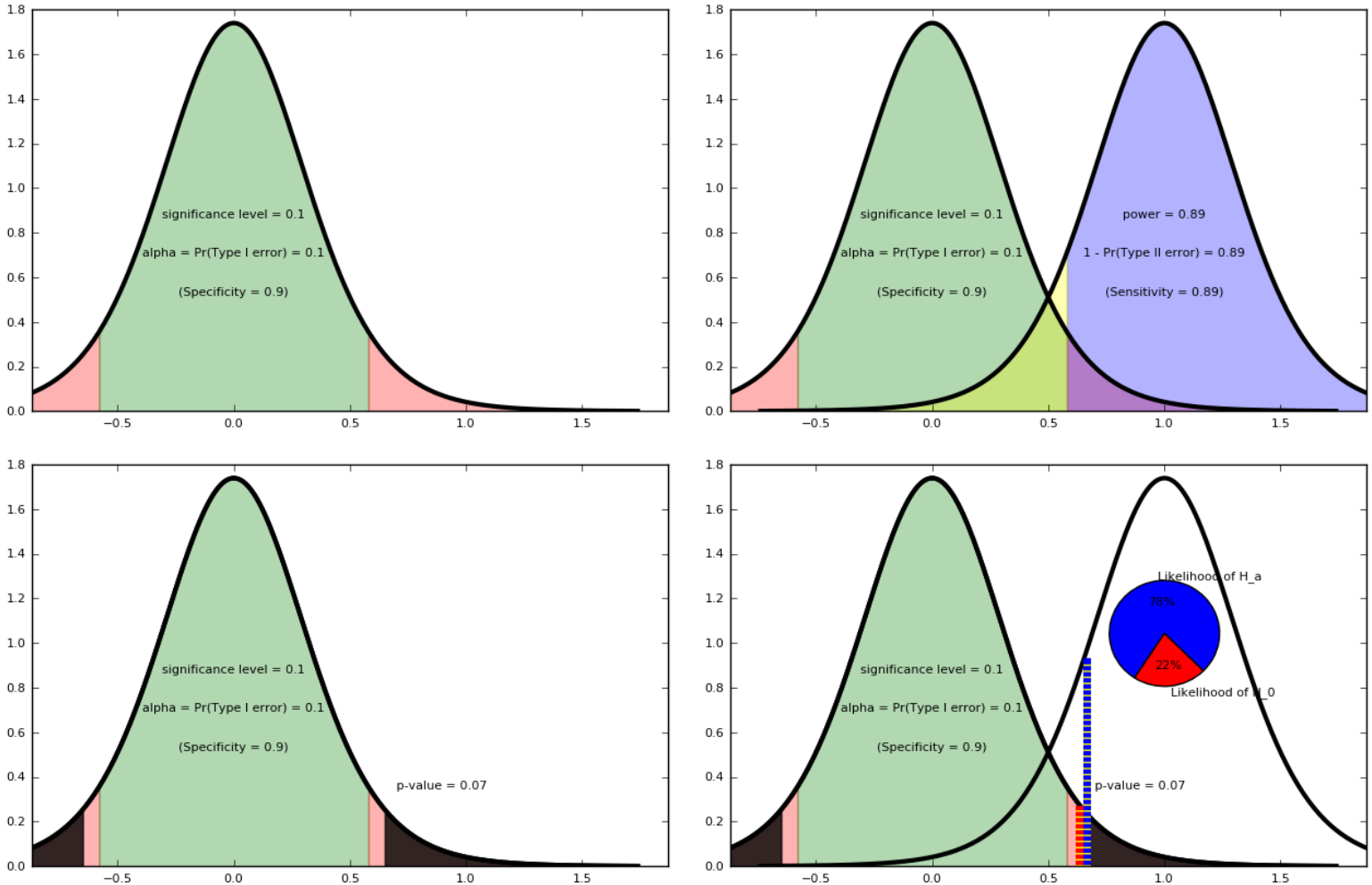
- What is Pr(H_0 Correctly Rejected)?

Hypothesis Testing



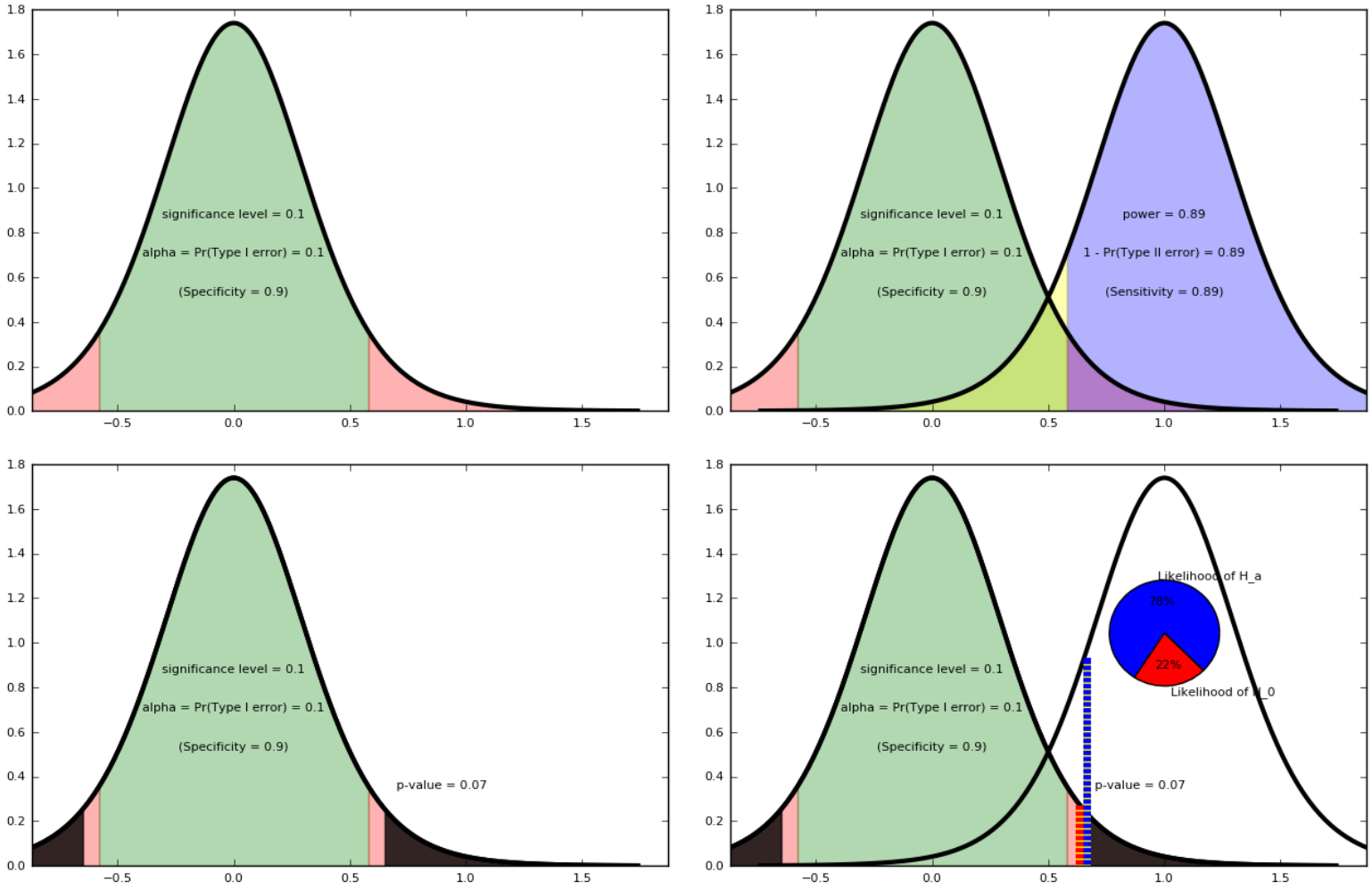
- What is Pr(H_0 Rejected | H_a True)?

Hypothesis Testing



- What is $\Pr(H_0 \text{ Accepted} \mid H_0 \text{ True})$?

Hypothesis Testing



- What is $\Pr(H_0 \text{ Accepted} \mid H_a \text{ True})$?

Hypothesis Testing *blunders I'll never forgive you for*

- ✗ A p-value *is not* the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

Hypothesis Testing *blunders I'll never forgive you for*

- X A p-value *is not* the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

- X A p-value *is not* the probability of incorrectly rejecting H_0
- ✓ Significance level α is the probability of wrongly rejecting H_0

Hypothesis Testing *blunders I'll never forgive you for*

- X A p-value *is not* the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

- X A p-value *is not* the probability of incorrectly rejecting H_0
- ✓ Significance level α is the probability of wrongly rejecting H_0

- X A p-value *is not anything else except*

$$\Pr(\textbf{as or more extreme than what you saw} | H_0)$$

Hypothesis Testing *blunders I'll never forgive you for*

X A p-value *is not* the probability H_0 is False

✓ H_0 is True, or it is not – there is no "sometimes/probability"

X A p-value *is not* the probability of incorrectly rejecting H_0

✓ Significance level α is the probability of wrongly rejecting H_0

X A p-value *is not anything else except*

$$\Pr(\textbf{as or more extreme than what you saw}|H_0)$$

✓ A p-value is *at all times ever only and EXACTLY ONLY*

$$\Pr(\textbf{as or more extreme than what you saw}|H_0)$$

The *Pivot*

If H_0 is true, then

$$.95 = \Pr\left(-1.96 < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < 1.96\right)$$

The *Pivot*

If H_0 is true, then

$$.95 = \Pr\left(-1.96 < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < 1.96\right)$$

$$= \Pr\left(-\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < -\mu_0 < -\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

The *Pivot*

If H_0 is true, then

$$
\begin{aligned}
.95 &= \Pr\left(-1.96 < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \\
&= \Pr\left(-\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < -\mu_0 < -\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} > \mu_0 > \bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right)
\end{aligned}
$$

The *Pivot*

If H_0 is true, then

$$
\begin{aligned}
.95 &= \Pr\left(-1.96 < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \\
&= \Pr\left(-\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < -\mu_0 < -\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} > \mu_0 > \bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < \mu_0 < \bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right)
\end{aligned}
$$

The *Pivot*

If H_0 is true, then

$$
\begin{aligned}
.95 &= \Pr\left(-1.96 < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \\
&= \Pr\left(-\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < -\mu_0 < -\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} > \mu_0 > \bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < \mu_0 < \bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right)
\end{aligned}
$$

- And that's a 95% Confidence Intervals

The *Pivot*

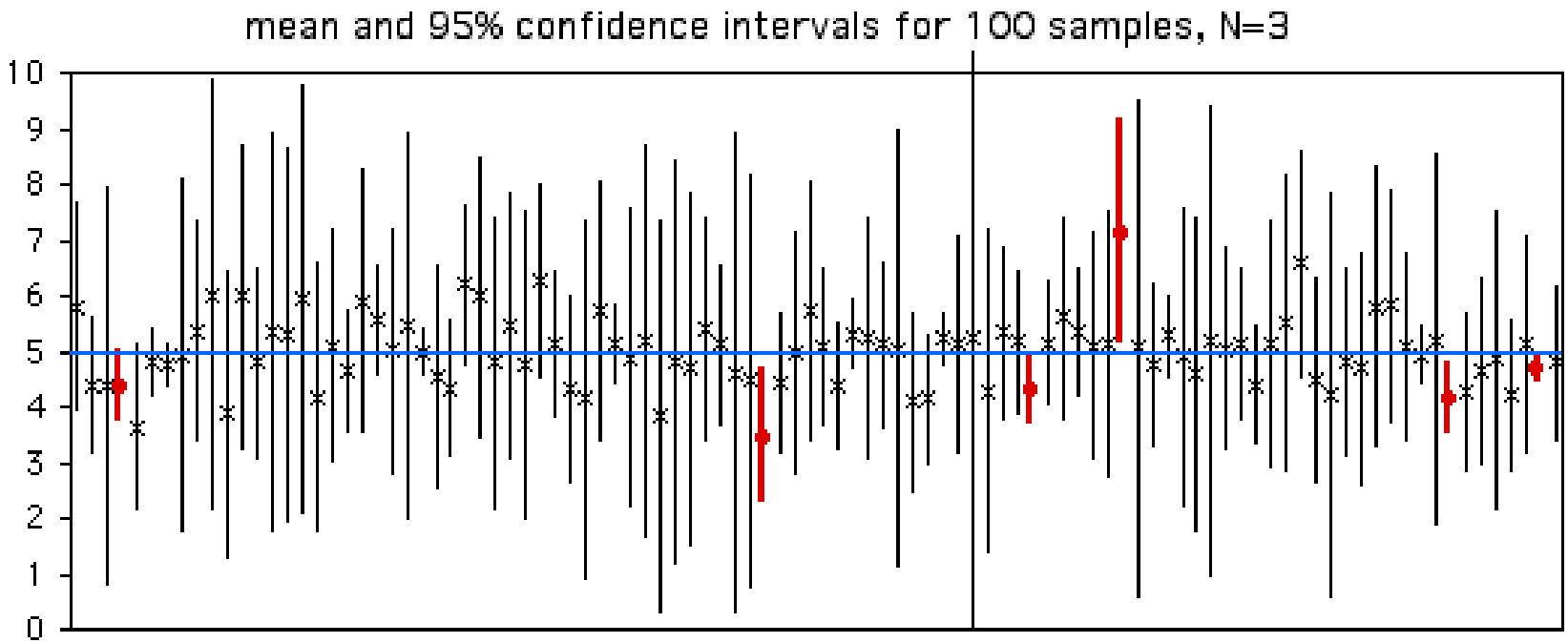
If H_0 is true, then

$$
\begin{aligned}
.95 &= \Pr\left(-1.96 < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \\
&= \Pr\left(-\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < -\mu_0 < -\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}} > \mu_0 > \bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right) \\
&= \Pr\left(\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} < \mu_0 < \bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right)
\end{aligned}
$$

- And that's a 95% Confidence Intervals

95% of the time it will "capture" μ_0
(under hypothetical repeated experimentation)

Confidence Intervals



Confidence Intervals and p-values *equivalence*

If the $100(1-\alpha)\%$ confidence interval *does not* contain μ_0 then H_0 will be rejected at the α-significance level

$$\Longleftrightarrow$$

If H_0 is rejected at the α-significance level then the $100(1-\alpha)\%$ confidence interval *will not* contain μ_0

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}} \left(\left| \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \right| > Z_{\alpha/2} \right)$$

The observed p-value under H_0 is

$$p = \Pr_Z \left(Z > \left| \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \right| \right)$$

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}}\left(\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}\right)$$

The observed p-value under H_0 is

$$p = \Pr_{Z}\left(Z > \left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right)$$

- If $p < \alpha$ then $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}$

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}}\left(\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}\right)$$

The observed p-value under H_0 is

$$p = \Pr_{Z}\left(Z > \left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right)$$

- If $p < \alpha$ then $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}$

$\Longrightarrow \mu_0 < \bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$ or $\mu_0 > \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}}\left(\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}\right)$$

The observed p-value under H_0 is

$$p = \Pr_{Z}\left(Z > \left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right)$$

- If $p < \alpha$ then $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}$

$\Longrightarrow \mu_0 < \bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$ or $\mu_0 > \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$

$\Longrightarrow \mu_0 \notin \left(\bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}}\left(\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}\right)$$

The observed p-value under H_0 is

$$p = \Pr_{Z}\left(Z > \left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right)$$

- If $p < \alpha$ then $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}$

$\Longrightarrow \mu_0 < \bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$ or $\mu_0 > \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$

$\Longrightarrow \mu_0 \notin \left(\bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$

the $100(1-\alpha)\%$ confidence interval *does not* contain μ_0

Multiple Testing

- Each time we do a hypothesis test [what?]

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

  And it guarantees a α *familly-wise error rate*

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

  And it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

  And it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

- An alternative is the *False Discovery Rate (FDR) q*

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

  And it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

- An alternative is the *False Discovery Rate (FDR) q*

  which for a set of tests (e.g., tests significant at the α-level)

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- Each time there's an α chance of being wrong

  So if we do N tests, we expect on average $\alpha \times N$ are wrong

  Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

  And it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

- An alternative is the *False Discovery Rate (FDR) q*

  which for a set of tests (e.g., tests significant at the α-level)

  is the proportion q of the tests called incorrectly ("FDR")