50.034 - Introduction to Probability and Statistics

Week 11 – Lecture 19

January–May Term, 2019

Lecture 19 Assigned Readings: 9.1

Outline of Lecture

- Introduction to hypothesis testing
- Null hypothesis and alternative hypothesis
- One-sided versus two-sided hypotheses
- Test statistic and rejection region
- Significance level
- Type I error and type II error
- Power of hypothesis tests
- p-value

Testing real-world hypotheses

Examples of real-world hypothesis testing:

1. Testing if more men than women suffer from nightmares.
2. Evaluating if the onset of the full moon has an effect on vehicle accident rate.
3. Determining whether a new medicine reduces headache.
4. Deciding whether a change in hand sanitizer brand would reduce the number of infections in a hospital.
5. Checking if the presence of designated smoking zones (i.e. those yellow boxes you sometimes see) help reduce smoking.

In each example, there is a specific **hypothesis** that can be tested.

- e.g. the claim "more men than women suffer from nightmares".

Hypothesis testing is a type of **statistical inference** to determine if the hypothesis fails to hold up after collecting experimental data.

- To make hypotheses precise and testable, we need to formulate suitable statistical models.

Note: "hypothesis" is singular, while "hypotheses" is plural.

The Main Idea of Hypothesis Testing

Consider a statistical model consisting of observable R.V.'s $X_1, \ldots, X_n$ that are conditionally iid given the parameter θ.

- Let Ω be the parameter space of θ, and consider any subset $\Omega_0 \subseteq \Omega$. Let H_0 denote the hypothesis "$\theta \in \Omega_0$".

Key Idea: H_0 is "tested" based on observed values for $X_1, \ldots, X_n$.

There are only two possible conclusions for testing H_0:

- **Reject** H_0, because observed values give sufficient statistical evidence to conclude that H_0 is not true.
- **Do not reject** H_0, because observed values are insufficient to conclude that H_0 is not true.

Question: If we do not reject H_0, does that mean $\theta \notin \Omega_0$?

Answer: NO!!

Analogy with Criminal Trial

Hypothesis testing is similar to a criminal trial.

Consider the following hypothesis

$$H_0 : \text{The defendant is innocent.}$$

- A defendant is considered not guilty as long as his or her guilt is not proven.
- The prosecutor tries to prove the guilt of the defendant. The defendant is convicted only when there is enough evidence.
- "*failure to reject* H_0" in this case does not imply innocence, but just means that there is **not enough evidence to convict.**
- So the jury does not *accept* H_0, but *fails to reject* H_0.

Similarly, in hypothesis testing, hypotheses can be rejected, but they CANNOT be accepted.

- Think of "do not reject H_0" as being inconclusive.
- Perhaps H_0 is true, but you do not know for sure.

Null hypothesis and alternative hypothesis

In hypothesis testing, the specific hypothesis H_0 to be tested is called the null hypothesis.

*[It is called "null hypothesis" because we want to **nullify** the null hypothesis.]*

Consider a statistical model consisting of observable R.V.'s $X_1, \ldots, X_n$ that are conditionally iid given the parameter θ.

- Let Ω be the parameter space of θ, and let H_0 be the null hypothesis "$\theta \in \Omega_0$", where $\Omega_0 \subseteq \Omega$ is some given subset.
- Let Ω_1 be the complement set of Ω_0 in Ω. The hypothesis H_1 defined by "$\theta \in \Omega_1$" is called the alternative hypothesis.
- **Note:** Either H_0 is true, or H_1 is true.

If we reject H_0, then we are asserting that there is sufficient statistical evidence to conclude that H_0 is not true, which means there is sufficient statistical evidence to conclude that H_1 is true.

- **Important Note:** If we **do not reject** the null hypothesis H_0, then there is insufficient evidence to conclude whether H_0 is false, which also means there is insufficent evidence to conclude whether the alternative hypothesis H_1 is true.

Example 1

Scenario: Your friend passes you a coin. He thinks that the coin is fair, but he is not sure. You guess that the coin is not fair.

You tell your friend that you can apply what you have learned from this course. You decide to use hypothesis testing.

- First, you set up a statistical model consisting of iid observable Bernoulli R.V.'s $X_1, \ldots, X_{100}$, representing 100 coin tosses, where $X_i = 1$ if the i-th toss is heads, and $X_i = 0$ otherwise. You assume the common parameter θ is an unknown constant.
- Next, you consider the following hypotheses:
  - H_0: $\theta = 0.5$ (null hypothesis);
  - H_1: $\theta \neq 0.5$ (alternative hypothesis).
- Based on the 100 coin tosses, you will decide whether to reject the null hypothesis (and conclude that the coin is NOT fair), or to not reject the null hypothesis (i.e. inconclusive).

Design of hypothesis tests

Example: Suppose there is a new pain relief medicine X. We want to determine if X relieves headaches faster than Panadol, and we wish to use hypothesis testing. There are two possible hypotheses we could choose as the null hypothesis H_0.

- "H_0: X relieves headaches faster than Panadol"?
- "H_0: X does not relieve headaches faster than Panadol"?

It is up to us to choose H_0, but which is "better"? It depends..

Important Consideration: Think of the criminal trial analogy.

- The alternative hypothesis is the actual hypothesis that we are interested in showing.
- Suppose we are the creators of X. We want to find enough evidence to "prove" that X relieves headaches faster than Panadol, which means we want to nullify H_0.
- If instead we work for Panadol, then we want to find enough evidence to "prove" that X does not relieve headaches faster than Panadol, which means we want to nullify H_0.

Simple and composite hypotheses

In the coin toss example, θ is a parameter that represents the success rate of tossing heads in your coin toss experiment.

- The parameter space Ω of θ is the interval $[0, 1]$.
- You chose $\Omega_0 = \{0.5\}$ for your null hypothesis "$H_0 : \theta \in \Omega_0$".
- Your chosen set Ω_0 contains just a single value 0.5.
  - Any hypothesis "$\theta \in \Omega_i$" is called simple if Ω_i contains exactly one value, and is called composite if Ω_i contains > 1 value.
  - Hence, your hypothesis "$H_0 : \theta = 0.5$" is a **simple hypothesis**.

Light Bulb Example with a composite null hypothesis:
Consider a statistical model consisting of iid observable exponential R.V.'s $X_1, \ldots, X_{100}$ with unknown constant parameter θ, where each X_i represents the lifespan (in hours) of the i-th light bulb.

- Assume that the parameter space of θ is the interval $(0, 1)$, i.e. a bulb lasts on average > 1 hour. Consider the null hypothesis

$$H_0 : \theta \leq \tfrac{1}{500}.$$

- This is a **composite hypothesis** and can be interpreted as the hypothesis "a light bulb lasts on average at least 500 hours".

One-sided hypothesis and one-tailed test

Suppose θ is a parameter with parameter space Ω.

- A null hypothesis is usually denoted by H_0.
- An alternative hypothesis is usually denoted by H_1.
  - If we want to define a hypothesis H_i that represents "$\theta \in \Omega_0$" for some subset $\Omega_0 \subseteq \Omega$, then we can simply write

$$H_i : \theta \in \Omega_0$$

- If Ω_0 is an interval of the form

  (k, ∞) or $[k, \infty)$ or $(-\infty, k)$ or $(-\infty, k]$,

  then we can write "$H_i : \theta \in \Omega_0$" simply as

  $H_i : \theta > k$ or $H_i : \theta \geq k$ or $H_i : \theta < k$ or $H_i : \theta \leq k$.

  In these 4 cases, we say that H_i is a one-sided hypothesis.
- Hypothesis testing done with a **one-sided** null hypothesis is usually called a one-tailed test.
  - If the null hypothesis is **one**-sided, then the alternative hypothesis consists of **one** interval, i.e. **one** "tail".

Two-sided hypothesis and two-tailed test

Same as before: Suppose θ is a parameter with parameter space Ω.

- If Ω_0 is an interval of the form

  $$(k_1, k_2) \quad \text{or} \quad [k_1, k_2) \quad \text{or} \quad (k_1, k_2] \quad \text{or} \quad [k_1, k_2],$$

  for some real numbers k_1, k_2, then we can write "$H_i : \theta \in \Omega_0$" simply as $H_i : k_1 < \theta < k_2$ or $H_i : k_1 \leq \theta < k_2$ or $H_i : k_1 < \theta \leq k_2$ or $H_i : k_1 \leq \theta \leq k_2$.

  In these 4 cases, we say that H_i is a two-sided hypothesis.
- Hypothesis testing done with a **two-sided** null hypothesis is usually called a two-tailed test.
  - If the null hypothesis is **two**-sided, then the alternative hypothesis consists of **two** intervals, i.e. **two** "tails".
- **Special Case:** If $\Omega_0 = [k, k] = \{k\}$ for some real number k (i.e. Ω_0 contains just a single value k), then we can write "$H_i : \theta \in \Omega_0$" simply as "$H_i : \theta = k$".

Intuition for hypothesis testing

Normal Distribution Example: Suppose we have a statistical model consisting of iid observable **normal** R.V.'s $X_1, \ldots, X_n$ with unknown constant mean θ and known variance σ^2. Let θ_0 be some given real number. Consider a two-tailed test with a null hypothesis

$$H_0 : \theta = \theta_0.$$

Question: Given $X_1 = x_1, \ldots, X_n = x_n$, how do we use these observed values to decide whether to reject or not reject H_0?

- By the law of large numbers, $\overline{X}_n \xrightarrow{p} \theta$, so we can compute $\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$ and check if $\overline{x}_n$ is "close to" θ_0.
- **Intuition:** If $\overline{x}_n$ is "too far from" μ_0, then $H_0 : \mu = \mu_0$ is very likely false, so we reject H_0.

Question: What about other non-normal distributions? What about other parameters?

- In this example, we are using the fact that θ is the mean. If instead $X_1, \ldots, X_n$ are exponential with parameter θ, then since $\mathbf{E}[X_i] = \frac{1}{\theta}$, we could then check if $\frac{1}{\overline{x}_n}$ is "close to" θ_0.

Test Statistic

More generally, given a null hypothesis $H_0 : \theta \in \Omega_0$, our decision on whether to reject H_0 or not reject H_0, will depend on the observed value of some statistic T.

- T is a statistic of the observable R.V.'s $X_1, \ldots, X_n$.
  - **Recall:** A statistic is a function of observable R.V.'s.
- e.g. if $X_1, \ldots, X_n$ are normal with mean θ, then our decision could be based on the observed value of the statistic $\overline{X}_n$.
- e.g. if $X_1, \ldots, X_n$ are exponential with parameter θ, then since $\mathbf{E}[X_i] = \frac{1}{\theta}$, our decision could then be based on the observed value of the statistic $\frac{1}{\overline{X}_n}$.

Definition: Suppose there is some set R of real numbers, such that the test procedure for H_0 is decided by "reject H_0" if the observed value of T is in R, and "do not reject H_0" if the observed value of T is not in R, then we say that T is a test statistic and that R is the rejection region of the test.

Steps for Hypothesis Testing

Model set-up: Let $X_1, \ldots, X_n$ be observable R.V.'s with unknown parameter θ. Let Ω be the parameter space of θ.

- Goal: Perform hypothesis testing on the parameter θ.

1. Specify some **null hypothesis** $H_0 : \theta \in \Omega_0$.
   - $\Omega_0 \subseteq \Omega$ is a subset chosen based on your specific application.
   - You wish to test whether the "true" value of θ is not in Ω_0.
2. Specify some **test statistic** $T = T(X_1, \ldots, X_n)$.
   - Your final decision will depend on the observed value of T.
3. Specify some **rejection region** $R \subseteq \mathbb{R}$.
   - This represents the region for where to reject H_0.
   - Note: R can be different from the complement of Ω_0.
4. Collect experimental evidence
   - Get observed values $X_1 = x_1, \ldots, X_n = x_n$.
5. Final decision: To reject or not to reject?
   - "Reject H_0" if $T(x_1, \ldots, x_n) \in R$.
   - "Do not reject H_0" if $T(x_1, \ldots, x_n) \notin R$.

The entire test procedure is collectively called a hypothesis test.

Example 2

Coin Toss Experiment: Toss a coin 10 times.

- Let $X_1, \ldots, X_{10}$ be iid Bernoulli R.V.'s with unknown constant parameter p, where $X_i = 1$ if the i-th coin toss is heads, and $X_i = 0$ otherwise.
- Consider the **null hypothesis** $H_0 : p = 0.5$.
  - In other words, we wish to test if the coin is biased or not.
  - Our guess is that the coin is biased, and we want to gather sufficient evidence to nullify the null hypothesis H_0.
- Let the **test statistic** be $T = X_1 + \cdots + X_{10}$.
  - T represents the total number of heads obtained (in 10 tosses).
  - For a fair coin, it is not surprising to get 4 or 6 heads.
  - Suppose we agree that getting ≤ 3 heads or ≥ 7 heads is unlikely if the coin is fair.
- Set the **rejection region** as $R = \{0, 1, 2, 3\} \cup \{7, 8, 9, 10\}$.
  - i.e. we reject H_0 if the observed value of T is not $4, 5$ or 6.

Some observations:

- The test statistic T does not have to be the sample mean!
- We can decide what the rejection region R is.

Significance level

Suppose we are given a null hypothesis $H_0 : \theta \in \Omega_0$.

Question: How do we decide on the rejection region R?

- We want to choose R so that if indeed the "true" value of θ is contained in Ω_0, then there is a "very low probability" that the observed value of the test statistic T would be contained in R.
- Equivalently, we want to choose R so that if H_0 is true, then there is a "very low probability" that H_0 will be rejected.
  - If H_0 is true but we reject H_0, then we have made an error.
  - Such an error is called a type I error.

To determine what we consider "very low probability", we need to state **thresholds** that we are willing to accept.

- We need to specify a threshold α_0 for the maximum tolerable probability that a **type I error** occurs.
- This threshold α_0 is called the level of significance, or more simply, the significance level.
  - The significance level α_0 is a real number!
  - Commonly used significance levels are 0.1, 0.05 or 0.01.

Significance level in terms of the power function

Let $\mathcal{H}$ be a hypothesis test with the null hypothesis $H_0 : \theta \in \Omega_0$, where θ is a parameter with parameter space Ω.
Suppose T is the test statistic, and let R be the rejection region.

- The power function of $\mathcal{H}$ is a function $\pi : \Omega \to \mathbb{R}$ defined by

$$\pi(\omega) = \Pr(T \in R | \theta = \omega)$$

  for every possible value $\omega \in \Omega$.
  - **Interpretation:** $\pi(\omega)$ is the probability that we will **reject** the null hypothesis H_0, given that the "true" value of θ equals ω.
  - In the course textbook, the notation $\pi(\omega|\mathcal{H})$ is used instead.

Definition: We say that $\mathcal{H}$ a level α_0 test, or equivalently, that $\mathcal{H}$ has a significance level of α_0, if $\pi(\omega) \leq \alpha_0$ for all $\omega \in \Omega_0$.

- **Interpretation:** "$\mathcal{H}$ is a level α_0 test" is exactly the same as "the probability that a type I error occurs for $\mathcal{H}$ is at most α_0."
- **Note:** There are many possible significance levels for $\mathcal{H}$.
  - The smallest possible significance level for $\mathcal{H}$ is called the size of $\mathcal{H}$.

Type I error versus type II error

Let $H_0 : \theta \in \Omega_0$ be a null hypothesis.

- A type I error occurs if H_0 is **true** but we **reject** H_0.
- A type II error occurs if H_0 is **false** but we **do not reject** H_0.

Important Note: Type I errors and type II errors are different!

Example: Let H_0 be "Patient A requires vitamin C supplements."

- Type I error: Patient A has a vitamin C deficiency, but the doctor has failed to prescribe vitamin C supplements.
- Type II error: Patient A does not have a vitamin C deficiency, but the doctor has prescribed vitamin C supplements.

In this example, a type I error is "more serious" than a type II error.

- We are not too concerned if a type II error occurs, but we do want type I errors to occur only with a "very low probability".
  - Here, we assume there is no danger of a vitamin C overdose.

Question: Should we minimize the probability of type I errors, e.g. try to get 0 probability for type I errors?

Trivial way to get zero probability for type I errors

Example: Let H_0 be "Patient A requires vitamin C supplements."

- Type I error: Patient A has a vitamin C deficiency, but the doctor has failed to prescribe vitamin C supplements.
- Type II error: Patient A does not have a vitamin C deficiency, but the doctor has prescribed vitamin C supplements.

If the doctor wants to minimize the probability that type I errors occur but does not care about type II errors, then the "easiest" way is to just prescribe vitamin C supplements for ALL patients.

- Unfortunately, this means the doctor would always make type II errors whenever H_0 is false! (worst case scenario for type II errors)
- Analogously, if the doctor wants to minimize the probability that type II errors occur without caring about type I errors, then he/she could just never prescribe vitamin C supplements.

Conclusion: We want to find a "good" balance: Ideally, we want a "very low probability" that type I errors occur, while simultaneously also having a "not-too-high" probability that type II errors occur.

Power of hypothesis test

Let $\mathcal{H}$ be a hypothesis test that consists of the following:

$$H_0 : \theta \in \Omega_0 \quad \text{(null hypothesis)};$$
$$H_1 : \theta \in \Omega_1 \quad \text{(alternative hypothesis)};$$

where θ is a parameter with parameter space $\Omega = \Omega_0 \cup \Omega_1$.
Suppose T is the test statistic, and let R be the rejection region.

- **Recall:** The power function of $\mathcal{H}$ is the function

$$\pi(\omega) = \Pr(T \in R | \theta = \omega) \qquad (\text{for } \omega \in \Omega).$$

- $\mathcal{H}$ has a significance level of α_0, if $\pi(\omega) \leq \alpha_0$ for all $\omega \in \Omega_0$.

Definition: Let β_0 be a real number. We say that $\mathcal{H}$ has a power of β_0, if $\pi(\omega) \geq \beta_0$ for all $\omega \in \Omega_1$.

- **Note:** power is a real number, while the power function is a function. The power gives a lower bound on the possible values of the power function on Ω_1.
- **Interpretation:** "$\mathcal{H}$ has power β_0" is exactly the same as "the probability that a type II error occurs is at most $1 - \beta_0$".
  - Higher power implies lower probability that type II errors occur.

Example 3

Let $X_1, \ldots, X_{100}$ be iid normal observable R.V.'s with unknown mean μ and known variance 25.

Consider a hypothesis test $\mathcal{H}$ with null hypothesis $H_0 : \mu = 10$. Let T be the test statistic $|\overline{X}_{100} - 10|$, and let R be the rejection region $[c, \infty)$, where c is some constant to be determined.

Determine the value of c that maximizes the power of $\mathcal{H}$ at significance level 0.01.

Example 3 - Solution

The given **null hypothesis** is $H_0 : \mu = 10$.

If indeed H_0 is true, then $\mathbf{E}[\overline{X}_{100}] = 10$ and $\mathrm{var}(\overline{X}_{100}) = \frac{25}{100} = 0.25$, so $\frac{\overline{X}_{100}-10}{\sqrt{0.25}} = 2(\overline{X}_{100} - 10)$ is a standard normal R.V.

We are told that H_0 is rejected if $|\overline{X}_{100} - 10| \geq c$. Note that

$$\begin{aligned}
\Pr(|\overline{X}_{100} - 10| \geq c) &= 1 - \Pr(-c \leq \overline{X}_{100} - 10 < c) \\
&= 1 - \Pr(-2c \leq 2(\overline{X}_{100} - 10) < 2c) \\
&= 1 - \Big(\Phi(2c) - \Phi(-2c)\Big) \\
&= 1 - \Big(\Phi(2c) - (1 - \Phi(2c))\Big) \\
&= 2 - 2 \cdot \Phi(2c),
\end{aligned}$$

where $\Phi(z)$ denotes the standard normal cdf.

Example 3 - Solution (continued)

Previous slide: $\Pr(|\overline{X}_{100} - 10| \geq c) = 2 - 2 \cdot \Phi(2c)$.

We want a significance level of 0.05, which means we want $2 - 2 \cdot \Phi(2c) \leq 0.05$, or equivalently, $\Phi(2c) \geq 0.975$.

From the standard normal cdf table, the closest value of z satisfying $\Phi(z) = 0.975$ is $z = 1.96$. Thus, $c \geq \frac{1.96}{2} = 0.98$.

We also want to maximize the power of $\mathcal{H}$.

Note: As c gets smaller, the rejection region $[c, \infty)$ gets larger, and the probability that a type II error occurs gets smaller.

- Type II error: do not reject H_0 when H_0 is false.
- Consequence: As c gets smaller, the power of $\mathcal{H}$ gets larger.

To maximize the power of $\mathcal{H}$ while still having that $\mathcal{H}$ is a level 0.05 test, we need to find the smallest c that satisfies $c \geq 0.98$.

Therefore, the value of c should be 0.98.

Example 4

Let $\{X_1, \ldots, X_{10}\}$ be a random sample of normal observable R.V.'s with unknown mean μ and unknown variance σ^2. Let $\overline{X}_{10}$ be the sample mean, and let

$$s_{10}^2(X_1, \ldots, X_{10}) = \frac{1}{9}\sum_{i=1}^{10}(X_i - \overline{X}_{10})^2$$

be the unbiased sample variance.

Suppose $\mathcal{H}$ is a hypothesis test with null hypothesis $H_0 : \mu = 5$, test statistic

$$T = \left| \frac{\sqrt{10}(\overline{X}_{10} - 5)}{s_{10}(X_1, \ldots, X_{10})} \right|,$$

and rejection region $R = [c, \infty)$, where c is some constant to be determined. Find the value of c that maximizes the power of $\mathcal{H}$ at significance level 0.01.

Example 4 - Solution

The given **null hypothesis** is $H_0 : \mu = 5$.

If indeed H_0 is true, then $Z = \frac{\sqrt{10}(\overline{X}_{10}-5)}{s_{10}(X_1,\ldots,X_{10})}$ has the t-distribution with 9 degrees of freedom.

We are told that H_0 is rejected if $T = |Z| \geq c$. Note that

$$\begin{aligned}
\Pr(|Z| \geq c) &= 1 - \Pr(-c \leq Z < c) \\
&= 1 - (F(c) - F(-c)) \\
&= 1 - (F(c) - (1 - F(c))) \\
&= 2 - 2 \cdot F(c),
\end{aligned}$$

where $F(z)$ denotes the cdf of Z.

We want a significance level of 0.01, which means we want $2 - 2 \cdot F(c) \leq 0.01$, or equivalently, $F(c) \geq 0.995$.

Example 4 - Solution (continued)

From the table of values for t-distributions, $F(3.250) = 0.995$.
Hence $F(c) \geq 0.995$ implies $c \geq 3.250$.
To maximize the power of $\mathcal{H}$ at significance level 0.01, we need to find the smallest possible c satisfying $c \geq 3.250$, thus $c = 3.250$.

Table of the *t* Distribution

If X has a t distribution with m degrees of freedom, the table gives the value of x such that $\Pr(X \leq x) = p$.

m	$p = .55$	.60	.65	.70	.75	.80	.85	.90	.95	.975	.99	.995
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977

Reporting the results of your hypothesis test

Example 3 (again): Let $\{X_1, \ldots, X_{100}\}$ be a random sample of normal observable R.V.'s with unknown mean μ and known variance 25. Consider a hypothesis test $\mathcal{H}$ with null hypothesis $H_0 : \mu = 10$, test statistic $T = |\overline{X}_{100} - 10|$, and rejection region $R = [0.98, \infty)$. From Example 3, we know that $\mathcal{H}$ has a significance level of 0.05.

Suppose after gathering experimental data, we reject H_0 because the observed value of T exceeds 0.98.

- Merely reporting that we have rejected H_0 is not so useful.
  - Was the observed value of T only slight above 0.98
  - Or was the observed value of T much larger than 0.98?
- What if we had decided to use a level 0.01 test? Would we still reject H_0? What about other significance levels?

To make the report of our experimental data more useful, we could report the **smallest significance level for which H_0 would be rejected**, given our observed value of the test statistic.

p-value

Let $T = T(X_1, \ldots, X_n)$ be a fixed statistic of a random sample $\{X_1, \ldots, X_n\}$ of observable R.V.'s with unknown parameter θ.

Let $\mathcal{H} = \{\mathcal{H}_c\}_{c \in \mathbb{R}}$ be a collection of hypothesis tests, where each $\mathcal{H}_c$ represents the hypothesis test with null hypothesis $H_0 : \theta \in \Omega_0$, test statistic T, and rejection region $[c, \infty)$. [Note: Ω_0 is a fixed subset.]

- Let α_c be the **size** of each $\mathcal{H}_c$, i.e. α_c is the smallest possible significance level for $\mathcal{H}_c$. (Different values of c give different sizes.)

Definition: Given some observed values $X_1 = x_1, \ldots, X_n = x_n$, let $t = T(x_1, \ldots, x_n)$ be the corresponding observed value of T. Then as c varies over $\mathbb{R}$, the smallest possible size α_c for which H_0 will be rejected given the observed value t, is called the p-value of $\mathcal{H}$.

- **Note:** The p-value depends on the observed values $x_1, \ldots, x_n$.
- **Interpretation:** If α is the p-value of $\mathcal{H}$, then it means that our experimental data is sufficient evidence to reject the null hypothesis H_0, whenever the value of c is chosen such that $\mathcal{H}_c$ has a significance level $\alpha_c \geq \alpha$.

Common misuse and abuse of p-values

p-values are commonly used in experimental physics, life sciences, economics, finance, and various social sciences.

- p-values have also been misused in these disciplines!
  - Usually, there is nothing wrong with the reported experimental data. Instead, the conclusions and interpretations drawn from the data are sometimes wrong but still get published.

Common Mistakes: (e.g. after finding a p-value of $\alpha < 0.05$)

- Conclude that the alternative hypothesis H_1 is true.
  - We can only conclude that the probability of type I errors is not more than α. There is still uncertainty whether H_1 is true or not, but we can use α to quantify our uncertainty.
  - We can say there is sufficient statistical evidence to conclude that H_1 is true, or that the results are statistically significant. This is NOT the same as H_1 being actually true!
  - Thus, any small p-value obtained is "heuristic evidence".
- Conclude that the null hypothesis H_0 is true with probability α.
  - $\Pr(H_0 \text{ is true}) = \Pr(\theta \in \Omega_0)$ is not the same as the smallest size α_c such that $T(x_1, \ldots, x_n) \geq c$ (given observed $x_1, \ldots, x_n$).

Common misuse and abuse of p-values (continued)

With the recent push by various scientists to raise awareness on the correct usage of p-values, it is now less likely (although still possible) to find published research with wrong usage of p-values. Unfortunately, even with correct usage, p-values can still be abused!

Common Abuse: (even with the correct usage of p-values)

- p-values depend heavily on the experimental design.
  - e.g. depends heavily on the specific choice of null hypothesis or test statistic.
  - Given the same experimental data, we can always modify our null hypothesis or test statistic to get different conclusions.
  - Modifying the null hypothesis or test statistic until a desired conclusion is obtained is a type of p-value hacking or p-hacking. (There are other types of p-hacking.)
- **Note:** Common significance levels used are 0.1, 0.05, 0.01.
  - These values are arbitrarily set. There is no "special meaning" when such significance levels are chosen for hypothesis testing.
  - Because of p-hacking, many reported p-values in various research articles are very close to one of these values.

Summary

- Introduction to hypothesis testing
- Null hypothesis and alternative hypothesis
- One-sided versus two-sided hypotheses
- Test statistic and rejection region
- Significance level
- Type I error and type II error
- Power of hypothesis tests
- p-value