9.07 Introduction to Statistics for Brain and Cognitive Sciences
Emery N. Brown

Lecture 12: Hypothesis Testing

I. Objectives

1. **Understand the hypothesis testing paradigm.**

2. **Understand how hypothesis testing procedures are constructed.**

3. **Understand how to do sample size calculations.**

4. **Understand the relation between hypothesis testing, confidence intervals, likelihood and Bayesian methods and their uses for inference purposes.**

II. The Hypothesis Testing Paradigm and One-Sample Tests

A. One-Sample Tests

To motivate the hypothesis testing paradigm we review first two problems. In both cases there is a single sample of data.

Example 3.2 (continued) Analysis of MEG Sensor Bias. Is there a noise bias in this SQUID sensor? Here we have a sample $x_1, ..., x_n$ where we model each $x_i \sim N(\mu, \sigma^2)$. We assume that μ is unknown and σ^2 is known. If there is bias, then $\mu \neq 0$, and $\mu > 0$ would suggest a positive bias where $\mu < 0$ would suggest a negative bias.

Example 2.1 (continued) Behavioral Learning. Has the animal learned the task? We have data $x_1, ..., x_n$ and we recall that our model is $x_i \sim B(n, p)$ where p is unknown. We define learning as performance greater than expected by chance. In particular, chance suggests $p = \frac{1}{2}$ (binary choice), learning suggests $p > \frac{1}{2}$, whereas impaired learning suggests $p < \frac{1}{2}$.

To define the hypothesis testing paradigm we state a series of definitions.

A **hypothesis** is the statement of a scientific question in terms of a proposed value of a parameter in a probability model. **Hypothesis testing** is a process of establishing proof by falsification. It has two essential components: a **null hypothesis** and an **alternative hypothesis**.

The **null hypothesis** is a stated value of the parameter which defines the hypothesis we want to falsify. It is usually stated as a single value although it can be composite. We denote it as H_0.

The **alternative hypothesis** is the hypothesis whose veracity we wish to establish. It is usually defined by a value or set of values of the parameter that are different from the one specified in the null hypothesis. We denote it as H_A.

To establish the result, we carry out a **test** in an attempt to reject the null hypothesis. The test is a procedure based on the observed data that allows us to choose between the null and alternative hypothesis.

Example 3.2 (continued) Analysis of MEG Sensor Bias. For this example the null hypothesis could be $H_0 : \mu = 0$ and alternative hypotheses could be

$$H_A : \mu > 0$$

or

$$H_A : \mu < 0.$$

Example 2.1 (continued) Behavioral Learning. Here the null hypothesis is $H_0 : p = \frac{1}{2}$ and the alternative hypotheses could be one-sided as either

$$H_A : p > \frac{1}{2}$$

or

$$H_A : p < \frac{1}{2}$$

or two-sided

$$H_A : p \neq \frac{1}{2}.$$

To investigate the hypothesis we require a **test statistic**. A test statistic is a statistic whose values will allow us to distinguish between the null and the alternative hypotheses.

Example 3.2 (continued) Analysis of MEG Sensor Bias. For this example we recall that $\bar{x} \sim N(\mu, \frac{\sigma^2}{n}).$ Hence, we can choose $\bar{x},$ because of its distributional properties. In general, choosing the statistic is an important issue. Here there are two cases:

Case i): If $\bar{x} >> 0$ or $\bar{x} << 0$ we conclude $\mu \neq 0.$ That is, we would be willing to conclude $\mu \neq 0$ if $|\bar{x}|$ is sufficiently large. Indeed, the larger $\bar{x}$ is in absolute value, the more likely we are to conclude $\mu \neq 0.$ If our data allow us to reach this conclusion, we say we **reject** the null hypothesis $H_0.$

Case ii) If $\bar{x}$ is close to $0,$ we say **we fail to reject** the null hypothesis $H_0.$ We do not say we accept the null hypothesis because, in this case, we do not reach a conclusion.

There are two types of errors we can commit in hypothesis testing. If we reject H_0 when it is true, this is an error. It is called a **Type 1 error**. The probability of the error is denoted as $\alpha.$ We write

$$\alpha = \Pr(\text{rejecting } H_0 \mid H_0 \text{ is true}).$$

In Example 3.2 we have

$$\alpha = \Pr(\bar{x} > c_\alpha \mid \mu = 0).$$

We choose α as small as possible. Typical values are $\alpha = 0.01$ and $\alpha = 0.05.$ What is c_α? To determine it, we compute

$$\Pr(\text{Type I error}) = \Pr(\text{rejecting } H_0 \mid H_0 \text{ true}) = \alpha.$$

We want α to be small

$$\begin{aligned} \alpha &= \Pr(\bar{x} > c_\alpha \mid \mu = \mu_0 = 0) \\ \alpha &= \Pr(\frac{n^{\frac{1}{2}}(\bar{x} - \mu_0)}{\sigma} > \frac{n^{\frac{1}{2}}(c_\alpha - \mu_0)}{\sigma}) \\ &= \Pr(\frac{n^{\frac{1}{2}}\bar{x} - \mu_0}{\sigma} > z_{1-\alpha}) \end{aligned}$$

where $z_{1-\alpha}$ is a quantile of the standard Gaussian. Hence, we have $c_\alpha = \mu_0 + z_{1-\alpha}\frac{\sigma}{n^{\frac{1}{2}}}.$ Some values of $z_{1-\alpha}$ for the corresponding values of α are

α	$z_{1-\alpha}$
0.05	1.645
0.025	1.96
0.01	2.325

The area to the right of $z_{1-\alpha}$ or c_α is the **critical region**. It has probability content of $\alpha.$ The value c_α is the **cut-off value**. This test based on the normal distribution is the **z-test**.

If H_0 is not true, i.e. H_A is true and we fail to reject $H_0,$ then this is also an error. It is termed a **Type II error** or β **error** and is defined as

$$\Pr(\text{Type II error}) = \Pr(\text{fail to reject } H_0 \mid H_A \text{ is true})$$

Assume H_A is true, i.e. $\mu = \mu_A > 0.$ Then we have

$$\begin{aligned} \beta &= \Pr(\bar{x} < c_\alpha \mid \mu = \mu_A) \\ &= \Pr(\frac{n^{\frac{1}{2}}(\bar{x} - \mu_A)}{\sigma} < \frac{n^{\frac{1}{2}}(c_\alpha - \mu_A)}{\sigma}) \\ &= \Pr(\frac{n^{\frac{1}{2}}(\bar{x} - \mu_A)}{\sigma} < z_\beta) \end{aligned}$$

We do not talk in terms of β but $1-\beta$ which is the **power of the test**. The power of the test is the probability of rejecting the null hypothesis when it is false. We compute it as

$$\begin{aligned} \text{power} = 1-\beta &= \Pr(\text{rejecting } H_0 \mid H_A \text{ is true}) \\ &= \Pr(\bar{x} > c_\alpha \mid \mu = \mu_A). \end{aligned}$$

Remark 12.1. You should not carry out a test if the power is not at least > 0.80 for important H_A.

Remark 12.2. Never report a negative result (failing to reject H_0) without reporting the power.

Remark 12.3. A statistical test is an “information assay.” As such, it is only as useful as it is powerful against an important alternative hypothesis.

If we reject H_0 we report the **p-value** which is the smallest value of α for which we can reject H_0. The **p-value** is also the probability of all events that are at least as rare as the observed statistic. It is the probability that we are making a mistake in rejecting the null hypothesis when the null hypothesis is true. An observed value of the test statistic has **statistical significance** if $p < \alpha$. **Statistical significance does not imply scientific significance.**

Example 3.2 (continued) Analysis of MEG Sensor Bias. In this example assume we have 500 observations $x_1, ..., x_n$ and the standard deviation is $\sigma = 1.1 \times 10^{-11} f$ Tesla . Suppose we wish to test $H_0 : \mu = 0$ against the alternative hypothesis $H_A : \mu > 0$ with $\alpha = 0.05$. If we have $\bar{x} = 0.11 \times 10^{-11} f$ Tesla then

$$z = \frac{n^{\frac{1}{2}}(\bar{x})}{\sigma} = \frac{(500)^{\frac{1}{2}}(0.11)}{1.1} = 2.25$$

From the Table of a standard normal distribution we have

$$z_{1-\alpha} = z_{0.95} = 1.645$$

or $\alpha = 0.05$ and we see $p = 0.0122.$ Therefore, we reject and conclude that there is a positive bias in the magnetic field around this recording sensor.

Example 2.1 (continued) Behavioral Learning. For this experiment if our hypotheses are

$$\begin{aligned} H_0 &: p = \frac{1}{2} \\ H_A &: p > \frac{1}{2} \end{aligned}$$

We have $n = 40$ trials and observe $k = 22$ and $\hat{p} = \frac{k}{40}$. Based on our Central Limit Theorem results in **Lecture 7**, we can analyze this question using the Gaussian approximation to the binomial provided $np > 5$ and $n(1-p) > 5.$ Because we have

$$np(1-p) = 40\left(\frac{1}{2}\right)\frac{1}{2} = 10 > 5$$

we can use the Gaussian approximation to the binomial to test H_0. Notice that if $np(1-p) > 5$ then it must be that $np > 5$ and $n(1-p) > 5$. If we take $\alpha = 0.05$, our test statistic is

$$z = \frac{n^{\frac{1}{2}}(\hat{p}-p)}{[p(1-p)]^{\frac{1}{2}}} = \frac{(40)^{\frac{1}{2}}(0.55-0.5)}{\left(\frac{1}{4}\right)^{\frac{1}{2}}} = \frac{6.32(0.05)}{\frac{1}{2}} = 0.632$$

$z_{1-\alpha} = 1.645$, hence $z < z_{1-\alpha}$ and we fail to reject H_0. We see that

$$c_\alpha = p + \frac{z_{1-\alpha}[p(1-p)]^{\frac{1}{2}}}{n^{\frac{1}{2}}} = \frac{1}{2} + \frac{1.645\left(\frac{1}{2}\right)}{6.32} = 0.63$$

What is the power of this test if the true probability of a correct response is $p_A = 0.72$?

$$\begin{aligned}
\text{power} = \Pr(\text{rejecting } H_0 \mid H_A = 0.72) &= \Pr(\hat{p} > 0.63 \mid p_A = 0.72) \\
&= \Pr(\frac{n^{\frac{1}{2}}(\hat{p}-p_A)}{[p_A(1-p_A)]^{\frac{1}{2}}} > \frac{n^{\frac{1}{2}}(c_\alpha - p_A)}{[p_A(1-p_A)]^{\frac{1}{2}}} \mid p_A = 0.72) \\
&= 1 - \Phi[\frac{(40)^{\frac{1}{2}}(0.63-0.72)}{[(0.18)(0.72)]^{\frac{1}{2}}} \\
&= 1 - \Phi[\frac{-(6.32)(0.09)}{(.36)}] \\
&= 1 - \Phi[-1.58] \\
&= 1 - 0.057 = 0.943.
\end{aligned}$$

Therefore, if the true propensity to respond correctly were $p_A = 0.72$ then there was a probability of 0.943 of rejecting the null hypothesis in this case.

B. One-Sample Test, Power and Sample Size Calculation

Given a null and a one-sided alternative hypothesis

$$\begin{aligned}
H_0 &: \mu = \mu_0 \\
H_A &: \mu = \mu_A > \mu_0
\end{aligned}$$

we compute the power as

$$\begin{aligned}
\text{Power} &= \Pr(\text{rejecting } H_0 \mid H_A \text{ is true}) \\
&= \Pr(\frac{n^{\frac{1}{2}}(\bar{x}-\mu_0)}{\sigma} > z_{1-\alpha} \mid \mu = \mu_A) \\
&= \Pr(\bar{x} > \mu_0 + \frac{z_{1-\alpha}\sigma}{n^{\frac{1}{2}}} \mid \mu = \mu_A) \\
&= \Pr(\bar{x} - \mu_A > \mu_0 - \mu_A + \frac{z_{1-\alpha}\sigma}{n^{\frac{1}{2}}} \mid \mu = \mu_A) \\
&= \Pr(\frac{n^{\frac{1}{2}}(\bar{x}-\mu_A)}{\sigma} > \frac{n^{\frac{1}{2}}(\mu_0-\mu_A)}{\sigma} + z_{1-\alpha} \mid \mu = \mu_A) \\
1-\Phi(\frac{n^{\frac{1}{2}}(\mu_0-\mu_A)}{\sigma} + z_{1-\alpha}) &= \Phi(-z_{1-\alpha} - \frac{n^{\frac{1}{2}}(\mu_0-\mu_A)}{\sigma}) \\
&= \Phi(z_\alpha + \frac{n^{\frac{1}{2}}(\mu_A-\mu_0)}{\sigma})
\end{aligned}$$

because $-z_{1-\alpha} = z_\alpha$. Similarly, if we have

$$\begin{aligned}
&H_0 : \mu = \mu_0 \\
&H_A : \mu = \mu_A < \mu_0
\end{aligned}$$

then it is easy to show that

$$\text{power} = \Phi[\frac{n^{\frac{1}{2}}(\mu_0-\mu_A)}{\sigma} + z_\alpha]$$

In general for a one-sided alternative

$$\text{power} = \Phi[\frac{n^{\frac{1}{2}} \mid \mu_0-\mu_A \mid}{\sigma} + z_\alpha]$$

We use these formulae to derive expressions for sample size calculations. Notice that

$$\begin{aligned}
\text{power} &= \Phi(\frac{n^{\frac{1}{2}} \mid \mu_0-\mu_A \mid}{\sigma} + z_\alpha) \\
\text{power} &= 1-\beta = \Phi(\frac{n^{\frac{1}{2}} \mid \mu_0-\mu_A \mid}{\sigma} + z_\alpha)
\end{aligned}$$

If we apply Φ^{-1} to the left and right hand sides of the equation above we get

$$\begin{aligned}
\Phi^{-1}(1-\beta) &= \Phi^{-1}[\Phi(\frac{n^{\frac{1}{2}} \mid \mu_A-\mu_0 \mid}{\sigma} + z_\alpha)] \\
z_{1-\beta} &= \frac{n^{\frac{1}{2}} \mid \mu_A-\mu_0 \mid}{\sigma} + z_\alpha \\
z_{1-\beta} - z_\alpha &= \frac{n^{\frac{1}{2}} \mid \mu_A-\mu_0 \mid}{\sigma}
\end{aligned}$$

Or since $-z_{\alpha} = z_{1-\alpha}$ we obtain the sample size formula

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha})^2}{\Delta^2}$$

where $\Delta = | \mu_0 - \mu_A |$.

Example 3.2 (continued) Analysis of MEG Sensor Bias. How many measurements should Steve Stufflebeam make daily to be at least 80% sure that if there is a positive drift of $0.1 \times 10^{-11} f$ Tesla he can detect it with $\alpha = 0.05$? To answer this question, we apply our sample size formula with $z_{0.95} = 1.645,\ z_{0.80} = 0.84,\ \sigma = 1.1 \times 10^{-1} f$ Tesla and we obtain

$$n = \frac{(1.1)^2 (1.645 + 0.84)^2}{(0.1)^2}$$
$$n = \frac{1.21 \times 6.18}{0.01}$$
$$n = 748$$

Therefore, in our problem studied above, we should have taken around 750 measurements instead of 500.

Remark 12.4. The ratio $\frac{\sigma^2}{\Delta^2}$ is like the inverse of a signal-to-noise ratio. As we want a smaller Type I error $(z_{1-\alpha})$, and/or more power n increases. Similarly, n increases with σ^2 and decreases with Δ^2.

III. One-Sample Two-Sided Tests
If the alternative hypothesis is two-sided then we need to construct a two-sided test.

A. Two-Sided Tests
Example 3.2 (continued) Analysis of MEG Sensor Bias. Suppose our null and alternative hypotheses for this problem are respectively

$$H_0 : \mu = 0$$
$$H_A : \mu \neq 0.$$

This alternative hypothesis implies that we would reject H_0 for either a positive bias or negative bias. Under H_0 we have $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$. We would therefore reject H_0 if $\bar{x} >> \mu_0 = 0$ or if $\bar{x} << \mu_0 = 0$. We take as our test statistic $\bar{x}$ and we will reject H_0 if $| \bar{x} | >> 0$. Pick α and take $\alpha = \alpha_1 + \alpha_2$. Since we do not favor $\mu > 0$ more than $\mu < 0$, we take $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$. To reject H_0, we consider

$$
\begin{aligned}
\Pr(|\bar{x}| > c_{\alpha/2}) &= \Pr(\bar{x} > c_{\alpha/2} \text{ or } \bar{x} < -c_{\alpha/2}) \\
&= \Pr\left(\frac{n^{\frac{1}{2}}(\bar{x}-\mu_0)}{\sigma} > \frac{n^{\frac{1}{2}}(c_{\alpha/2}-\mu_0)}{\sigma} \text{ or } \frac{n^{\frac{1}{2}}(\bar{x}-\mu_0)}{\sigma} < -\frac{n^{\frac{1}{2}}(c_{\alpha/2}-\mu_0)}{\sigma}\right) \\
&= \Pr(z > z_{1-\alpha/2} \text{ or } z < -z_{1-\alpha/2}) \\
&= \Pr(|z| > z_{1-\alpha/2})
\end{aligned}
$$

This is a **two-sided test because we reject** H_0 for either very large positive or negative values of the test statistic. We reject H_0 for $|z| > z_{1-\alpha/2}$ or equivalently, we reject H_0 if

$$|\bar{x}| > c_{\alpha/2} = \mu_0 + \frac{\sigma}{n^{\frac{1}{2}}} z_{1-\frac{\alpha}{2}}$$

Examples of α and $z_{1-\frac{\alpha}{2}}$ are

α	$z_{1-\frac{\alpha}{2}}$
0.10	1.645
0.05	1.96
0.01	2.58

Example 3.2 (continued) Analysis of MEG Sensor Bias. We consider

$$
\begin{aligned}
H_0 &: \mu = 0 \\
H_A &: \mu \neq 0
\end{aligned}
$$

Suppose we pick $\alpha = 0.05,$ we assume $\sigma = 1.1$ and we compute $\bar{x} = -0.11,$ then we have

$$z = \frac{n^{\frac{1}{2}}(\bar{x})}{\sigma} = \frac{-(500)^{\frac{1}{2}}0.11}{1.1} = -2.25$$

and $z_{1-\alpha/2} = z_{0.975} = 1.96$. Because $-2.25 < -1.96,$ we have $z < -z_{1-\alpha/2}$ and hence, we reject H_0.

Example 2.1 (continued) Behavioral Learning. We can perform a similar analysis for the learning example

$$
\begin{aligned}
H_0 &: p_0 = \frac{1}{2} \\
H_A &: p \neq \frac{1}{2}
\end{aligned}
$$

This alternative implies either impaired learning or learning and would lead us to reject if $\hat{p} >> \frac{1}{2}$ or $\hat{p} << \frac{1}{2}$. We have that under the Gaussian approximation

$$z = \frac{n^{\frac{1}{2}}(\hat{p} - p_0)}{[p_0(1-p_0)]^{\frac{1}{2}}}$$

Hence, given α we reject H_0 if

$$z > z_{1-\alpha/2} \text{ or } z < z_{\alpha/2}$$

or equivalently if

$$\hat{p} > c_{\alpha/2} \text{ or } \hat{p} < -c_{\alpha/2}$$

where $c_{\alpha/2} = p_0 + [\frac{p_0(1-p_0)}{n}]^{\frac{1}{2}} z_{1-\alpha/2}$. Given $\alpha = 0.10$, we obtain $z_{1-\alpha/2} = z_{0.95} = 1.645$ and since $\hat{p} = \frac{k}{40} = \frac{22}{40}$, we have

$$z = \frac{n^{\frac{1}{2}}(\hat{p} - p_0)}{[p_0(1-p_0)]^{\frac{1}{2}}} = \frac{(40)^{\frac{1}{2}}(0.55 - 0.5)}{(\frac{1}{4})^{\frac{1}{2}}} = \frac{6.32(0.05)}{\frac{1}{2}} = 0.632$$

Because $z < z_{0.95}$ we fail to reject H_0.

B. Power for the Two-Sided Alternative

It is straightforward to show that for the mean of a Gaussian distribution with known variance if the null hypothesis is $H_0 : \mu = \mu_0$ versus the two-sided alternative $H_A : \mu \neq \mu_0$ the power of the two-sided test is defined as

$$\Pr(|\frac{n^{\frac{1}{2}}(\bar{x} - \mu_0)}{\sigma}| > z_{1-\alpha/2} \mid H_A \text{ is true}) = \Pr(\frac{n^{\frac{1}{2}}(\bar{x} - \mu_0)}{\sigma} > z_{1-\alpha/2} \text{ or } \frac{n^{\frac{1}{2}}(\bar{x} - \mu_0)}{\sigma} < z_{\alpha/2} \mid H_A \text{ is true})$$

This simplifies to

$$\text{power} = \Phi[-z_{1-\alpha/2} + \frac{n^{\frac{1}{2}}(\mu_0 - \mu_A)}{\sigma}] + \Phi[-z_{1-\alpha/2} + \frac{n^{\frac{1}{2}}(\mu_A - \mu_0)}{\sigma}].$$

The corresponding sample size formula is

$$n = \frac{\sigma^2}{\Delta^2}(z_{1-\alpha/2} + z_{1-\beta})^2.$$

where $\Delta = |\mu_A - \mu_0|$.

Example 3.2 (continued) Analysis of MEG Sensor Bias. If Steve wanted to worry about both positive and negative drift, then the previous sample size calculation becomes with $z_{1-\alpha/2} = z_{0.975} = 1.96,$

$$
\begin{aligned}
n &= \frac{(1.1)^2(1.96+0.84)^2}{(0.1)^2} \\
&= \frac{1.21(2.8)^2}{0.01} = \frac{1.21(7.84)}{0.01} \\
&= 949.
\end{aligned}
$$

Remark 12.5. Notice that $(z_{0.975}+z_{0.8})^2 \approx 8.$ Hence,

$$n \approx \frac{8}{SNR}$$

Remark 12.6. In **Homework Assignment 9** we will explore a similar formula for the binomial distribution.

C. Adjustments for the Gaussian Assumptions

1. One-Sample t-Test for Unknown σ^2. The $z-$test allows us to test hypotheses about the mean of a Gaussian distribution under the assumption that the variance is known. The t-test allows us to test the same hypotheses when the sample size is not large and the variance is unknown and must be estimated from the sample. The t-test was developed by Gossett in 1908 while he was working in the Guinness Brewery. Gossett wrote under the pseudonym of Student. For this reason it is still referred to as Student's *t*-test. The distribution was worked out later by R.A. Fisher.

Suppose $x_1,...,x_n$ is a random sample from a Gaussian probability model $N(\mu,\sigma^2)$ and we wish to test the null hypothesis $H_0: \mu = \mu_0$ against the alternative $H_A: \mu \neq \mu_0$.

Assume σ^2 is not known and n is not large, say $15 < n < 20.$ Therefore, as discussed in **Lecture 8**, we construct a t-test by estimating σ^2 with an unbiased estimate, and instead of a $z-$statistic we construct a t-statistic as

$$t = \frac{n^{\frac{1}{2}}(\overline{x}-\mu_0)}{s}$$

where

$$s^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

$t \sim t_{n-1},$ a t-distribution on $n-1$ degrees of freedom. Recall that we showed in the practice problems for the second In Class Examination that s^2 is an unbiased estimate of σ^2. Given $\alpha,$ to test $H_0: \mu = \mu_0$ against $H_A: \mu \neq \mu_0$ we reject H_0 for

$$|t| > t_{n-1,1-\alpha/2}$$

or equivalently if either

$$\bar{x} > \mu_0 + t_{n-1,1-\alpha/2}\frac{s}{n^{\frac{1}{2}}} \text{ or } \bar{x} < \mu_0 - t_{n-1,1-\alpha/2}\frac{s}{n^{\frac{1}{2}}}$$

Example 12.1 Reaction Time Measurements. In a learning experiment, along with the correct and incorrect responses, we record the reaction times which are the times it takes the animal to execute the task. In a previous study, once it had been determined that an animal had learned a task, it was found that the average reaction time was 10 seconds. On the 14 trials after the animal learned the task by the behavioral criteria, the average reaction time was 8.25 seconds. The sample standard deviation was 2.1. Is this animal's performance different from that previously reported? We have

$$H_0 : \mu = 10.0$$
$$H_A : \mu \neq 10.0$$

$$t = \frac{n^{\frac{1}{2}}(\bar{x}-\mu)}{s} = \frac{(14)^{\frac{1}{2}}(8.25-10)}{(2.1)} = \frac{(3.74)(-1.75)}{(2.1)} = -2.26$$

Now it follows from the Table of the t-distribution that $t_{13,0.975} = 2.16$. Because $|t| > t_{13,0.975}$ we reject H_0.

2. Binomial Exact Method

If $np_0(1-p_0) < 5,$ we cannot use the Gaussian approximation to the binomial to tests hypotheses about the binomial proportion. In this case, we base the test on the exact binomial probabilities. We have $x \sim B(n,p)$ and we observe k successes, and we take $\hat{p} = \frac{k}{n}.$ The p-value depends on whether $\hat{p} \leq p_0$ or $\hat{p} > p_0$. If $\hat{p} \leq p_0$, then

$$\frac{p-value}{2} = \Pr(\leq k \text{ successes in } n \text{ trials} \mid H_0)$$
$$= \sum_{j=0}^{k} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

If $\hat{p} > p_0$, then

$$\frac{p-value}{2} = \Pr(\geq k \text{ successes in } n \text{ trials} \mid H_0)$$
$$= \sum_{j=k}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

Example 12.2. A New Learning Experiment. Assume that we execute the learning experiment with 20 trials and there is a $\frac{1}{3}$ probability of a correct response by chance. Suppose $k = 12$ and $\hat{p} = \frac{12}{20} = \frac{3}{5}.$ We want to test $H_0 : p = \frac{1}{3}$ against $H_A : p \neq \frac{1}{3}.$ We see that

$$np_0(1-p_0) = 20\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{40}{9} = 4\frac{4}{9} < 5$$

We have $\frac{12}{20} > \frac{1}{3}$ and hence $\hat{p} > p_0$ and we compute the p-value as

$$\frac{p}{2} = \Pr(x \ge 12) = \sum_{j=12}^{20} \binom{20}{j}\left(\frac{1}{3}\right)^{j}\left(\frac{2}{3}\right)^{20-j}$$
$$= 0.01286$$

or equivalently

$$p = 2(0.01286) = 0.02572.$$

Therefore, we reject H_0 and conclude that the animal is not performing at chance and most likely has learned.

Remark 12.7. An important topic that we have not considered is non-parametric tests. Each of the main parametric tests we considered, i.e., the z-test and t-test has a non-parametric analog. It is important to use these nonparametric tests when the sample size is small and the Gaussian assumption on which most of the tests are based is not valid (Rosner, 2006).

Remark 12.8. The requirement of the Gaussian assumption for most the standard hypothesis testing paradigms is very limiting. For this reason, we have spent more time in the course learning about principles of modern statistical analysis such as likelihood methods, the bootstrap and other Monte Carlo methods, Bayesian methods and, as we shall see in **Lecture 16**, the generalized linear model.

D. The Relation Among Confidence Intervals, Hypothesis Tests and p- Values

The three statistics we have considered thus far for two-sided hypothesis tests can be used to construct $100 \times (1-\alpha)$ confidence intervals. These are

z- test (Gaussian mean and variance known)

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{n^{\frac{1}{2}}}$$

z- test (Binomial proportion)

$$\hat{p} \pm z_{1-\alpha/2}\left[\frac{\hat{p}(1-\hat{p})}{n}\right]^{\frac{1}{2}}$$

t- test (Gaussian mean and variance unknown)

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{n^{\frac{1}{2}}}$$

We can reject H_0 with $\alpha = 0.05$ if the value of the parameter under the null hypothesis is not in the $100\%(1-\alpha)$ confidence interval. In this way, a confidence interval provides a hypothesis test. A similar construction of a confidence bound can be used to carry out a one-sided test. **Most importantly, the confidence interval tells us the reasonable range for the parameter that can be inferred from the data analysis.** Confidence intervals report the analysis results on the physical scale on which the problem takes place. The p-value only tells us how likely the observed statistic would be under H_0. In this way, **the hypothesis test simply provides a mechanism for making a decision. Confidence intervals are always more informative than p-values.** Realizing this in the early 80s, the *New England Journal of Medicine* set out a recommendation obliging that all statistical results be reported in terms of confidence intervals. This is now the standard for publication of research articles in that journal. Indeed, p-values alone mean nothing!

Example 3.2 (continued) Analysis of MEG Sensor Bias. To illustrate this point, we note that in this problem, the 95% confidence interval is

$$0.11 \pm 1.96 \frac{1.1}{(500)^{\frac{1}{2}}}$$

$$0.11 \pm 0.096$$

$$[0.014, \quad 0.206]$$

and we reject H_0 because 0 is not in the confidence interval. The p-value was < 0.05 which essentially tells us nothing about the magnitude of the bias in the sensor.

Example 2.1 (continued) Behavioral Learning. In this problem, the 95% confidence interval is

$$\hat{p} \pm 1.96 \frac{\left(\frac{1}{4}\right)^{\frac{1}{2}}}{40^{\frac{1}{2}}}$$

$$0.55 \pm 1.96 \frac{\frac{1}{2}}{6.32}$$

$$0.55 \pm 0.158$$

$$[0.392, 0.708].$$

Hence, the p-value is > 0.05. We fail to reject the null hypothesis. The confidence interval not only makes it clear that the null hypothesis is not rejected, it also shows what the reasonable range of uncertainty is for the animal's propensity to learn.

Remark 12.9. One of the true values of the hypothesis-testing paradigm is sample size calculations.

IV. Two-Sample Tests

A. Two-Sample $t-$Test

Example 12.3. Reaction Time Analysis in a Learning Experiment. Suppose we have the reaction times on trial 50 of the 9 rats from the treatment group and 13 rats from the control group we studied in **Homework Assignment 8-9**. The mean reaction time for the treatment group was 15.5 sec with a standard deviation of 3.25 sec and the mean reaction time for the control group was 11.5 sec with a standard deviation of 3.1 sec. What can be said about the underlying mean differences in reaction times between the two groups?

Assume we have

$$x_i^T \sim N(\mu_T, \sigma_T^2) \qquad i = 1, ..., n_T$$

$$x_j^c \sim N(\mu_c, \sigma_c^2) \qquad j = 1, ..., n_c$$

Take

$$H_0 : \mu_T = \mu_c$$

$$H_A : \mu_T \neq \mu_c$$

We have

$$\begin{aligned}\bar{x}_T &= 15.5\,\text{sec}\\ s_T &= 3.25\,\text{sec}\\ \bar{x}_c &= 11.5\,\text{sec}\\ s_c &= 3.1\,\text{sec}\end{aligned}$$

Under H_0 we have $E(\bar{x}_T - \bar{x}_c) = E(\bar{x}_T) - E(\bar{x}_c) = \mu_T - \mu_c = 0$ and hence,

$$\bar{x}_T \sim N(\mu_T, \frac{\sigma_T^2}{n_T})$$

$$\bar{x}_c \sim N(\mu_c, \frac{\sigma_c^2}{n_c})$$

where $n_T = 9$ and $n_c = 13.$

Now under the independence assumption of the two samples

$$Var(\bar{x}_T - \bar{x}_c) = Var(\bar{x}_T) + Var(\bar{x}_c) = \frac{\sigma_T^2}{n_T} + \frac{\sigma_c^2}{n_c}$$

Hence, under H_0 and the assumption that $\sigma_T^2 = \sigma_c^2 = \sigma^2$,

$$\overline{x}_T - \overline{x}_c \sim N(0, \sigma^2(\frac{1}{n_T} + \frac{1}{n_c})).$$

If σ^2 were known, then we would have

$$\frac{\overline{x}_T - \overline{x}_c}{\sigma(\frac{1}{n_T} + \frac{1}{n_c})^{\frac{1}{2}}} \sim N(0,1)$$

and we could base our hypothesis test on this z-statistic. Since σ^2 is unknown, let us consider the estimate of σ^2 defined by

$$s^2 = \frac{(n_T - 1)s_T^2 + (n_c - 1)s_c^2}{n_T + n_c - 2}$$

where

$$s_T^2 = (n_T - 1)^{-1} \sum_{j=1}^{n_T} (x_j^T - \overline{x}_T)^2$$

$$s_c^2 = (n_c - 1)^{-1} \sum_{i=1}^{n_c} (x_i^c - \overline{x}_c)^2.$$

Notice that if we assume that $\sigma_T^2 = \sigma_c^2 = \sigma^2$ then

$$E(s^2) = \frac{(n_T - 1)E(s_T^2) + (n_c - 1)E(s_c^2)}{n_T + n_c - 2} = \frac{(n_T - 1)\sigma^2 + (n_c - 1)\sigma^2}{n_T + n_c - 2} = \sigma^2,$$

and s^2 is an unbiased estimate of σ^2.

Given α we can test H_0 with the following test statistic

$$t = \frac{\overline{x}_T - \overline{x}_c}{s(\frac{1}{n_T} + \frac{1}{n_c})^{\frac{1}{2}}}$$

termed the two-sample t-statistic with equal variance with $n = n_T + n_c - 2$ degrees of freedom. We reject H_0 at level α

$$| t | > t_{n,1-\alpha/2}$$

Example 12.3 (continued). For this problem we have

$$\begin{aligned} s^2 &= \frac{(n_T - 1)s_T^2 + (n_c - 1)s_c^2}{n_T + n_c - 2} \\ &= \frac{8(3.25)^2 + (12)(3.1)^2}{20} \\ &= \frac{8(10.56) + 12(9.61)}{20} \\ &= \frac{84.48 + 115.32}{20} \\ &\frac{199.8}{20} = 9.999 \end{aligned}$$

and hence,

$$t = \frac{15.5 - 11.5}{[10 \times (\frac{1}{9} + \frac{1}{13})]^{\frac{1}{2}}} = \frac{4}{(1.88)^{\frac{1}{2}}} = 2.92$$

From the Table of the t-distribution we have $t_{20,0.975} = 2.086.$ Since $t > t_{20,0.975}$ we reject H_0 and conclude that the mean reaction times of the two groups are different. The longer average reaction time for the treatment group suggests that learning may be impaired in that group.

B. Confidence Interval for the True Difference in the Means

Because our test statistic follows a t-distribution, we can construct a $100\%(1-\alpha)$ confidence interval for the true difference in the measure as follows

$$\overline{x}_T - \overline{x}_c \pm t_{n,1-\alpha/2} s \left(\frac{1}{n_T} + \frac{1}{n_C} \right)^{\frac{1}{2}}.$$

Example 12.3 (continued) Reaction Time Analysis in a Behavioral Experiment. If we apply this formula to the data from **Example 12.3,** we obtain with $\alpha = 0.05$ a 95% confidence interval for the true mean difference of

$$4 \pm 2.086 \times [10 \times \left(\frac{1}{9} + \frac{1}{13} \right)]^{\frac{1}{2}}$$

$$4 \pm 2.086 \times (1.37)$$

$$4 \pm 2.86$$

which is

$$[1.14 \ \ 6.86].$$

The interval does not contain zero as expected based on our hypothesis test. More importantly, we see that not only is it unlikely that the true mean difference is $0,$ but that the difference could be as small as 1.14 sec or as large as 6.86 sec.

Remark 12.10. If we cannot assume that the variances in the two samples are equal but unknown, then we have to devise an alternative t-statistic that takes account of the unknown and estimated variances. It can be shown that an appropriate t-statistic in this case is

$$t = \frac{\overline{x}_T - \overline{x}_c}{\left(\frac{s_T^2}{n_T} + \frac{s_c^2}{n_c}\right)^{\frac{1}{2}}}$$

where the number of degrees of freedom is

$$d' = \frac{(\frac{s_T^2}{n_T} + \frac{s_c^2}{n_c})^2}{\left(\frac{s_T^2}{n_T}\right)^2 (n_T - 1)^{-1} + \left(\frac{s_c^2}{n_c}\right)^2 (n_c - 1)^{-1}}$$

This statistic is the Satterthwaite approximation to the degrees of freedom for a t-statistic with unknown and unequal variance (Rosner, 2006).

C. Two-Sample Test for Binomial Proportions

Example 2.1 (continued) Behavioral Learning Experiment. Let us assume that on Day 1, the rat had $k_1 = 22$ correct responses and on Day 2 we had $k_2 = 15$ correct responses. Day 1 is out of 40 trials and Day 2 is out of 20 trials. What can we say about the difference in performance between the two days?

We could treat the results from Day 1 as “truth” and compare Day 2 to Day 1. A more appropriate way to proceed is to treat the data from both days as if they were observed with uncertainty. For this we assume

$$x_j^1 \sim B(n_1, p_1) \qquad j = 1, ..., n_1$$

$$x_j^2 \sim B(n_2, p_2) \qquad j = 1, ..., n_2.$$

Take

$$H_0 : p_1 = p_2$$

$$H_A : p_1 \neq p_2$$

Our test statistic can be derived from the estimates of p_1 and p_2

$$\hat{p}_1 = \frac{k_1}{n_1}$$

$$\hat{p}_2 = \frac{k_2}{n_2}$$

Under H_0 and using the Gaussian approximation to the binomial

$$\hat{p}_1 \approx N(p, \frac{p(1-p)}{n_1})$$

$$\hat{p}_2 \approx N(p, \frac{p(1-p)}{n_1})$$

and if we assume the samples are independent we have the approximate z-statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{[\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)]^{\frac{1}{2}}} \approx N(0,1)$$

where, since p is unknown, we estimate it as $\hat{p}$ defined as

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2}$$

Hence, given a level α we have the approximate z-statistic is

$$z = \frac{|\hat{p}_1 - \hat{p}_2|}{[\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)]^{\frac{1}{2}}}$$

We reject H_0: if $z > z_{1-\alpha/2}$ and the approximate p-value is $p = 2[1-\Phi(z)]$.

An alternative form of z that includes the continuity correction (**Lecture 8**) to make the Gaussian approximation to the binomial more accurate is defined as

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{[\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)]^{\frac{1}{2}}}.$$

Example 2.1 (continued) Behavioral Learning Experiment. For this problem, we have

$$\hat{p} = \frac{k_1 + k_2}{n_1 + n_2} = \frac{22+15}{60} = \frac{37}{60} = 0.6167$$
$$1 - \hat{p} = 0.3833$$

or

$$
\begin{aligned}
z &= \frac{|0.55-0.75|}{[(0.6167)(0.3833)\left(\frac{1}{40}+\frac{1}{20}\right)]^{\frac{1}{2}}} \\
&= \frac{0.20}{[0.2363\overline{8}\left(\frac{3}{40}\right)]^{\frac{1}{2}}} = \frac{0.20}{(0.01773)^{\frac{1}{2}}} = \frac{0.20}{0.133} = 1.50.
\end{aligned}
$$

Because $z_{1-\alpha/2} = z_{0.975} = 1.96$ we have $z_{0.975} > |z|$ so we fail to reject the null hypothesis of no difference in performance. Here the p-value is

$$p = 2[1-\Phi(1.50)] = 2[0.0668] = 0.1336.$$

Similarly, the 95% confidence interval is

$$
\begin{aligned}
&\hat{p}_1 - \hat{p}_2 \pm z_{0.975}[\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)]^{\frac{1}{2}} \\
&-0.20 \pm 1.96(0.133) \\
&-0.20 \pm 0.26
\end{aligned}
$$

or

$$[-0.46, 0.06].$$

As expected, the 95% confidence interval includes zero which explains why we fail to reject H_0. The confidence interval suggests that there may be evidence of improved performance on the second day relative to the first since most of the interval includes negative values of the difference.

The value of the continuity correction is $\left(\frac{1}{2n_1}+\frac{1}{2n_2}\right) = \left(\frac{1}{80}+\frac{1}{40}\right) = 0.0375.$ This changes z from 1.50 to $z = 1.4625$ and the corrected p-value is $0.1442.$

Remark 12.11. The two sample binomial data can also be analyzed as a 2×2 contingency table

	Correct	Incorrect	Total
Day 1	k_1	$n_1 - k_1$	n_1
Day 2	k_2	$n_2 - k_2$	n_2
	$k_1 + k_2$	$n_1 + n_2 - (k_1 + k_2)$	$n_1 + n_2$

A 2×2 contingency table is a table consisting of two rows cross-classified by two columns. The contingency table analysis uses a test statistic based on a chi-squared distribution with one degree of freedom and gives the same result as the z-test. This is to be expected since we showed in **Lecture 4** that the square of a standard Gaussian random variable is a chi-squared random variable with one degree of freedom. We will study this in **Homework Assignment 9.**

Remark 12.12. Use of the Gaussian approximation to the binomial to construct this z-test is valid provided $n_1 p_1 \geq 5,\ n_2 p_2 \geq 5,\ n_1(1-p_1) \geq 5$ and $n_2(1-p_2) \geq 5.$ This corresponds to the condition that the analysis of the contingency table using a chi-squared statistic is valid if the expected number of observations per cell is at least 5.

Remark 12.13. This problem could easily be analyzed using a Bayesian analysis in which p_1 and p_2 had uniform priors. We could then use **Algorithm 10.1** to compare the posterior densities of p_1 and p_2.

Remark 12.14. We can perform a likelihood analysis and compare the overlap in the likelihoods. We could alternatively construct $1-\alpha$ confidence intervals for p_1 and p_2 separately using the likelihood theory and see if they overlap.

Remark 12.15. All the tests we have discussed here can be derived from the likelihood theory we presented by the likelihood ratio procedure. A detailed discussion of this approach is beyond the scope of this course. (See DeGroot and Schervish, 2002; Rice, 2007).

V. Summary

Hypothesis testing is a key part of classical statistics. It emphasizes procedures based primarily on the Gaussian distribution and Gaussian approximations. The hypothesis testing paradigm is very useful for prospective planning of studies using sample size formulae. **Confidence intervals are always more informative than p-values.** Confidence intervals report the analysis results on the physical scale on which the problem takes place. Many of the classical problems in hypothesis testing can now be carried out in a more informative way using more modern approaches such as Monte Carlo methods, bootstrapping, and Bayesian methods.

Acknowledgments

I am grateful to Julie Scott for technical assistance in preparing this lecture and to Jim Mutch for careful proofreading and comments.

References

DeGroot MH, Schervish MJ. *Probability and Statistics*, 3rd edition. Boston, MA: Addison Wesley, 2002.

Rice JA. *Mathematical Statistics and Data Analysis*, 3rd edition. Boston, MA, 2007.

Rosner B. *Fundamentals of Biostatistics*, 6th edition. Boston, MA: Duxbury Press, 2006.