Hypothesis Testing

Recall that a *point estimate* of some parameter is its most plausible value, in the light of some observed data.

Similarly, an *interval estimate* is a range of reasonably plausible values.

Sometimes, a particular value of the parameter is of interest, and we want to decide how plausible it is, again in the light of some observed data.
Example

A foundry making 16GB flash memory chips has historically had a 3% loss rate to process flaws. New equipment has a greater throughput, but a test batch of 250 chips contains 12 with flaws, a 4.8% rate.

Was that just a chance effect, or is the new equipment more prone to flaws?
The statistical framework: \( X \) is the number of flawed chips, and we assume that flaws arise independently, so \( X \sim \text{Bin}(n, p) \).

The simplest explanation is that nothing changed, that is \( p = p_0 = .03 \). We call this the *null hypothesis* and denote it \( H_0 \).

\[ H_0 : p = p_0. \]

The alternative is that something did change, and we’re especially concerned that it’s worse. This *alternative hypothesis* is denoted \( H_a \).

\[ H_a : p > p_0. \]
Note: neither $H_0$ nor $H_a$ allows the possibility that the new equipment is *better*: $p < p_0$.

We should really express the null hypothesis as “the new equipment is no worse than the current equipment”, and then $H_0$ becomes

$$H_0 : p \leq p_0.$$ 

Now all possibilities are covered.

In other cases, we may be interested in changes in either direction:

$$H_0 : p = p_0$$

$$H_a : p \neq p_0.$$
We now ask: 

*If $H_0$ were true, what is the chance of seeing as many as 12 flaws in 250 trials?*

And the answer is .076 when $p = p_0 = .03$, although less when $p < p_0$.

So finding 12 or more flawed chips is not especially unlikely under the null hypothesis, and we would not regard it as strong evidence that $H_0$ is false.

In any situation, we can carry out a similar calculation: the probability of observing something as extreme as what actually happened, if the null hypothesis were true.
The result is called the $P$-value, and is written $P = .076$, for example.

By convention, $P < .05$ is regarded as “evidence against $H_0$”, and $P < .01$ is regarded as “strong evidence”.

A $P$-value $.1 > P \geq .05$ might be called “weak evidence”.
Test Procedures

Sometimes we need to make a *decision* about the null hypothesis, not just weigh the evidence against it; e.g., whether to accept the new equipment, or ask the supplier to fix it.

We must decide whether or not to *reject* the null hypothesis.

Note: a null hypothesis is usually unlikely to be *exactly* true, so we do not speak of *accepting* it, only failing to reject it.

Think of it as a *working* hypothesis, which we use as an approximation until it’s shown to be false.
Test procedure

To carry out a hypothesis test, we need:

- A *test statistic*, such as the count $X$ of faulty chips.
- Usually, a cutoff point, or *critical value*, to identify values of the test statistic for which we reject $H_0$, such as $X > 12$.
- Formally, a *rejection region*: the set of values of the test statistic for which we reject $H_0$, such as $\{13, 14, \ldots \}$. 
Errors

Making a decision about a null hypothesis has the possibility of two kinds of error:

**Type I error:** Rejecting the null hypothesis when it is true;

**Type II error:** Failing to reject the null hypothesis when it is false.

Error Probabilities

Conventionally, the probabilities of Type I and Type II errors are denoted $\alpha$ and $\beta$, respectively.
In cases like the chip foundry, where the hypotheses are

\[ H_0 : p \leq 0.03 \]
\[ H_a : p > 0.03 \]

both \( \alpha \) and \( \beta \) depend on \( p \).

If the rule is to reject \( H_0 \) when \( X > 12 \),

\[ \alpha(p) = P(X > 12) = \sum_{x=13}^{250} b(x; 250, p), \quad p \leq 0.03 \]

and

\[ \beta(p) = P(X \leq 12) = \sum_{x=0}^{12} b(x; 250, p), \quad p > 0.03 \]
Significance level

We usually ignore the dependence of $\alpha(p)$ on $p$ by looking only at the worst case.

The *significance level* of the test, also denoted $\alpha$, is the worst Type I error probability.

In the chip foundry example, this is

$$\alpha = \max_{0<p\leq.03} \alpha(p)$$

and this is easily shown to be $\alpha(.03) = .0402$. 
Power

The dependence of $\beta(p)$ on $p$ cannot be handled as simply: if $p$ is just a little greater than .03,

$$
\beta(p) = \sum_{x=0}^{12} b(x; 250, p) \approx \sum_{x=0}^{12} b(x; 250, .03) = 1 - \alpha = .9598
$$

but, for larger $p$, $\beta(p)$ is more reasonable.

For example, $\beta(.05) = .5175$, and $\beta(.10) = .0021$.

We usually focus on

$$
\text{Power}(p) = P(\text{Reject } H_0) \text{ as a function of } p
= 1 - \beta(p).
$$
The power curve:

```R
plot(function(p) 1 - pbinom(12, 250, p), from = .03, to = .10,
     xlab = "p", ylab = "Power", ylim = c(0, 1))
title("Power curve")
abline(h = 1 - pbinom(12, 250, .03), col = "blue")
```
Tests About a Population Mean

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean $\mu$.

To decide how plausible is a particular value $\mu_0$, it is natural to see how far the sample mean $\bar{x}$ is from $\mu_0$.

If $\bar{x}$ is close to $\mu_0$, that value seems quite plausible, but not otherwise.

Suppose that we are interested in deviations in either direction:

\[ H_0 : \mu = \mu_0 \]
\[ H_a : \mu \neq \mu_0. \]
For example, 36 water samples taken downstream from the discharge of a water treatment facility showed barium concentrations with $\bar{x} = 10.87$ and $s = 13.31$ mg/L, respectively, whereas the upstream concentration was 5.32 mg/L.

The (estimated) standard error of $\bar{X}$ is

$$\frac{13.31}{\sqrt{36}} = 2.22$$

so the observed downstream mean is

$$\frac{10.87 - 5.32}{2.22} = 2.50$$

standard errors higher than upstream.
The natural test statistic is

$$|T| = \frac{|\bar{X} - \mu_0|}{\text{standard error of } \bar{X}}$$

where $T$ has observed value

$$t = \frac{\bar{x} - \mu_0}{\text{standard error of } \bar{X}}.$$

In the example,

$$t = \frac{10.87 - 5.32}{2.22} = 2.50$$

as we calculated earlier.
To test $H_0$, we need to calculate the $P$-value

$$P(|T| \geq |t| \text{ when } H_0 \text{ is true}).$$

We can do this in various cases:

- $X_1, X_2, \ldots, X_n$ normally distributed, $\sigma$ known: $T \sim N(0, 1)$;
- $X_1, X_2, \ldots, X_n$ normally distributed, $\sigma$ unknown but estimated by $s$: $T \sim \text{Student's } t$;
- $n$ large, $\sigma$ known or estimated by $s$: $T \approx N(0, 1)$. 
In the example, we could use the large sample size 36 to justify using the normal distribution, and calculate

\[ P(|T| \geq 2.50) \approx 1 - \Phi(2.50) + \Phi(-2.50) = .012 \]

Alternatively, we could guess that the individual measurements are normally distributed, and use the \( t \)-distribution with \( n - 1 = 35 \) degrees of freedom:

\[ P(|T| \geq 2.50) = 1 - F_{35}(2.50) + F_{35}(-2.50) = .017 \]

Either way, \( P < .05 \) and the \( P \)-value is close to .01, so we have evidence against \( H_0 \), if not strong evidence.
Test Procedure

If we must make a decision, we need a *rejection region*.

Typically, we first choose the significance level $\alpha$, most commonly .05.

The critical value is then either $z_{\alpha/2}$ or $t_{\alpha/2, n-1}$, depending on which assumptions we are making.

For instance, in the normal case,

$$P(|T| \geq z_{\alpha/2}) = 1 - \Phi(z_{\alpha/2}) + \Phi(-z_{\alpha/2}) = \alpha/2 + \alpha/2 = \alpha.$$  

Then we reject $H_0$ whenever $|t| \geq$ critical value.
One-sided Hypotheses and Tests

Most often, when a null hypothesis is loosely stated as \( H_0 : \mu = \mu_0 \), the appropriate alternative hypothesis is the two-sided \( H_a : \mu \neq \mu_0 \).

Sometimes, deviations in the different directions have such different implications that the alternative should be one-sided, and then the correct \( H_0 \) takes the opposite side.
Example: Workplace Compliance

Regulations: worker’s exposure to benzene must be less than 1 ppm.

The onus is on the employer to show that the limit has not been breached.

\( H_0 \) is that the workplace is not in compliance, \( H_0 : \mu \geq 1 \), and the alternative is \( H_a : \mu < 1 \).

Clearly we reject \( H_0 \) only if \( \bar{x} \) is sufficiently below \( \mu_0 \).

For significance level \( \alpha \), reject \( H_0 \) if

\[
T \leq -z_\alpha \text{ or } -t_{\alpha,n-1}.
\]

Note that \( \alpha \) is not divided by 2 in this case.
Power and Sample Size
Consider the two-sided case:

\[ H_0 : \mu = \mu_0 \]
\[ H_a : \mu \neq \mu_0 \]

and suppose, unrealistically, that \( \sigma \) is known.

The test statistic is \( |T| \), where

\[ T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}. \]
Under $H_0$, $T \sim N(0, 1)$, so we reject $H_0$ when $|T| \geq z_{\alpha/2}$.

Under $H_a$, write $\mu = \mu_0 - \delta$, with $\delta \neq 0$. Then

$$T \sim N\left(\frac{-\delta}{\sigma/\sqrt{n}}, 1\right)$$

so

$$Z = T + \frac{\delta}{\sigma/\sqrt{n}} \sim N(0, 1).$$
So

\[ P(|T| \geq z_{\alpha/2}) = P\left( Z \leq \frac{\delta}{\sigma/\sqrt{n}} - z_{\alpha/2} \right) \]

\[ + P\left( Z \geq \frac{\delta}{\sigma/\sqrt{n}} + z_{\alpha/2} \right) \]

\[ = \Phi\left( \frac{\delta}{\sigma/\sqrt{n}} - z_{\alpha/2} \right) \]

\[ + 1 - \Phi\left( \frac{\delta}{\sigma/\sqrt{n}} + z_{\alpha/2} \right) . \]
This is complicated, but can at least be graphed.

For example, suppose that \( n = 36 \) and \( \sigma = 13.31 \):

\[
\begin{align*}
n &\leftarrow 36 \\
sigma &\leftarrow 13.31 \\
\text{Power} &\leftarrow \text{function}(\delta, n, \sigma, \alpha = .05) \{ \\
& \quad \text{se} \leftarrow \sigma / \sqrt{n} \\
& \quad z \leftarrow \text{qnorm}(1 - \alpha/2) \\
& \quad \text{pnorm}(\delta/\text{se} - z) + 1 - \text{pnorm}(\delta/\text{se} + z) \\
\}\end{align*}
\]

\[
\text{curve(Power(x, n, sigma), from = -10, to = 10,} \\
& \quad \text{xlab = expression(delta), ylab = "Power")} \\
\text{abline(h = .05, col = "green")} \\
\text{title(main = "Power curve")}
\]
Suppose that a difference of $\delta = \pm 5$ is considered to be substantial.

From the graph, the power is around 0.6; $\text{Power}(5, n, \sigma)$ gives the value 0.616.

Often we want to have at least an 80% chance of detecting a substantial difference.

Trial-and-error shows that $\text{Power}(5, 56, \sigma)$ is just over 0.8, so we would need a sample size of at least $n = 56$ to achieve this.
In the more realistic case where \( \sigma \) is unknown, critical values are from the \( t \)-distribution, and power calculations involve the noncentral \( t \)-distribution (see also Table A.17).

Deciding on a sample size by specifying the length of a confidence interval (usually 95%) is far simpler than using power curves.
Binomial Probability

The first example dealt with a hypothesis about the probability parameter \( p \) in the binomial distribution \( \text{Bin}(n, p) \).

If \( n \) is large and neither \( np \) nor \( n(1 - p) \) is small, we use the fact that, under the null hypothesis \( H_0 : p = p_0 \), \( \hat{p} \approx N[p_0, p_0(1 - p_0)/n] \).

So

\[
T = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \approx N(0, 1)
\]

We can use \( T \) to test \( H_0 \) against the two-sided alternative \( H_a : p \neq p_0 \), or one of the one-sided alternatives \( H_a : p > p_0 \) or \( H_a : p < p_0 \), using \( z \)-based critical values.
Small samples

When we cannot use the normal approximation, or if we just choose not to, we can calculate exact binomial probabilities to get the $P$-value in a one-sided test.

For the binomial distribution, or any other discrete distribution such as the Poisson, the $P$-value changes in jumps as the observed $x$ changes.

We cannot in general find a critical value that gives exactly a specified significance level like .05.
It is clear that

\[ |t| \geq \text{critical value} \]

if and only if

\[ P(|T| \geq |t|) \leq \alpha. \]

That is, we reject \( H_0 \) at significance level \( \alpha \) if and only if the \( P \)-value \( \leq \alpha \).

So calculating the \( P \)-value both:

- weighs the evidence against \( H_0 \);
- allows the formal test to be carried out at any chosen significance level \( \alpha \).