

Fixed Level Testing, Power

Dr.Mamatha.H.R

Professor

Department of Computer Science and
Engineering

PES University

Bangalore

Course material created using various
Internet resources and text book

Fixed-Level Testing

- A hypothesis test measures the plausibility of the null hypothesis by producing a P -value.
- The smaller the P -value, the less plausible the null.
- there is no scientifically valid dividing line between plausibility and implausibility, so it is impossible to specify a “correct” P -value below which we should reject H_0 .
- When possible, it is best simply to report the P -value, and not to make a firm decision whether or not to reject.

- For example, if a sample of parts is drawn from a shipment and checked for defects, a decision must be made whether to accept or to return the shipment.

If a decision is going to be made on the basis of a hypothesis test, there is no choice but to pick a cutoff point for the P -value.

When this is done, the test is referred to as a **Fixed-level** test.

- Fixed-level testing is just like the hypothesis testing we have been discussing so far, except that a firm rule is set ahead of time for rejecting the null hypothesis.

A value α , where $0 < \alpha < 1$, is chosen. Then the P -value is computed.

If $P \leq \alpha$, the null hypothesis is rejected and the alternate hypothesis is taken as truth.

If $P > \alpha$, then the null hypothesis is considered to be plausible.

The value of α is called the **significance level**, or, more simply, the **level**, of the test.

if a test results in a P -value less than or equal to α , we say that the null hypothesis is rejected at level α (or $100\alpha\%$), or that the result is statistically significant at level α (or $100\alpha\%$).

a common choice for α is 0.05.

Summary

To conduct a fixed-level test:

- Choose a number α , where $0 < \alpha < 1$. This is called the significance level, or the level, of the test.
- Compute the P -value in the usual way.
- If $P \leq \alpha$, reject H_0 . If $P > \alpha$, do not reject H_0 .

- The mean wear in a sample of 45 steel balls was $\bar{X} = 673.2\mu\text{m}$, and the standard deviation was $s = 14.9\mu\text{m}$. Let μ denote the population mean wear. A test of $H_0 : \mu \geq 675$ versus $H_1 : \mu < 675$ yielded a P -value of 0.209. Can we reject H_0 at the 25% level? Can we reject H_0 at the 5% level?
- **Solution**
- The P -value of 0.209 is less than 0.25, so if we had chosen a significance level of $\alpha = 0.25$, We would reject H_0 .
- Thus we reject H_0 at the 25% level.
- Since $0.209 > 0.05$, we do not reject H_0 at the 5% level.

Critical Points and Rejection Regions

- In a fixed-level test, a **critical point** is a value of the test statistic that produces a P -value exactly equal to α .
- A critical point is a dividing line for the test statistic just as the significance level is a dividing line for the P -value.
- If the test statistic is on one side of the critical point, the P -value will be less than α , and H_0 will be rejected.
- If the test statistic is on the other side of the critical point, the P -value will be greater than α , and H_0 will not be rejected.
- The region on the side of the critical point that leads to rejection is called the **rejection region**.
- The critical point itself is also in the rejection region.

- A new concrete mix is being evaluated. The plan is to sample 100 concrete blocks made with the new mix, compute the sample mean compressive strength \bar{X} , and then test $H_0 : \mu \leq 1350$ versus $H_1 : \mu > 1350$, where the units are MPa. It is assumed from previous tests of this sort that the population standard deviation σ will be close to 70 MPa. Find the critical point and the rejection region if the test will be conducted at a significance level of 5%.

- reject H_0 if the P -value is less than or equal to 0.05.
- The P -value for this test will be the area to the right of the value of X .
- Therefore the P -value will be less than 0.05, and H_0 will be rejected, if the value of X is in the upper 5% of the null distribution.
- The rejection region therefore consists of the upper 5% of the null distribution.
- The critical point is the boundary of the upper 5%.

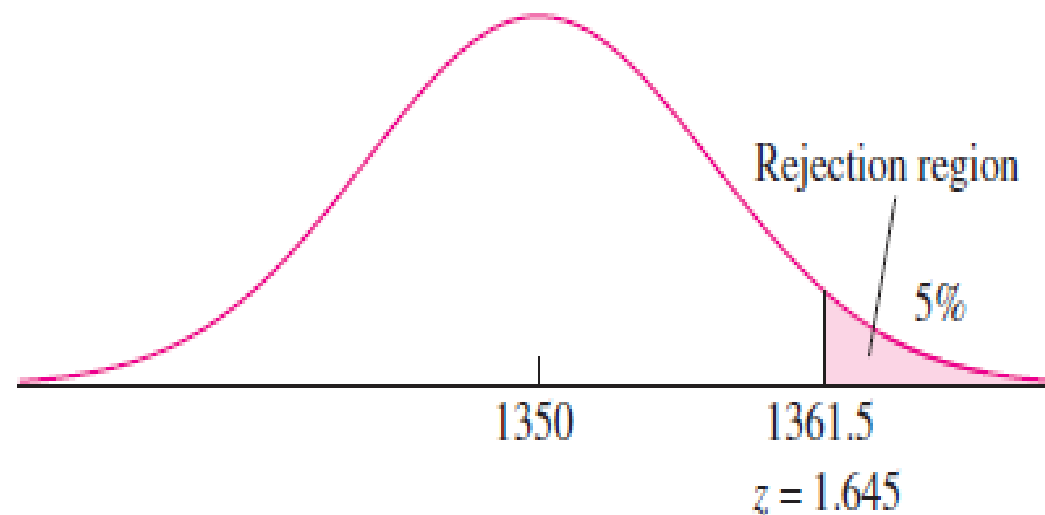


FIGURE 6.23 The rejection region for this one-tailed test consists of the upper 5% of the null distribution. The critical point is 1361.5, on the boundary of the rejection region.

- The null distribution is normal, and from the z table find that the z-score of the point that cuts off the upper 5% of the normal curve
- $z_{.05} = 1.645$.
- Therefore we can express the critical point as $z = 1.645$, and the rejection region as $z \geq 1.645$.
- express the critical point and rejection region in terms of X , by converting the z-score to the original units.
- The null distribution has mean $\mu = 1350$
- standard deviation $\sigma_X = \sigma/\sqrt{n} \approx 70/\sqrt{100} = 7$.
- Therefore the critical point can be expressed as
- $X = 1350 + (1.645)(7) = 1361.5$.
- The rejection region consists of all values of X greater than or equal to 1361.5.

- In a hypothesis test to determine whether a scale is in calibration, the null hypothesis is $H_0 : \mu = 1000$ and the null distribution of X is $N(1000, 0.262)$. Find the rejection region if the test will be conducted at a significance level of 5%.
- **Solution**
- Since this is a two-tailed test, the rejection region is contained in both tails of the null distribution.
- Specifically, H_0 will be rejected if X is in either the upper or the lower 2.5% of the null distribution

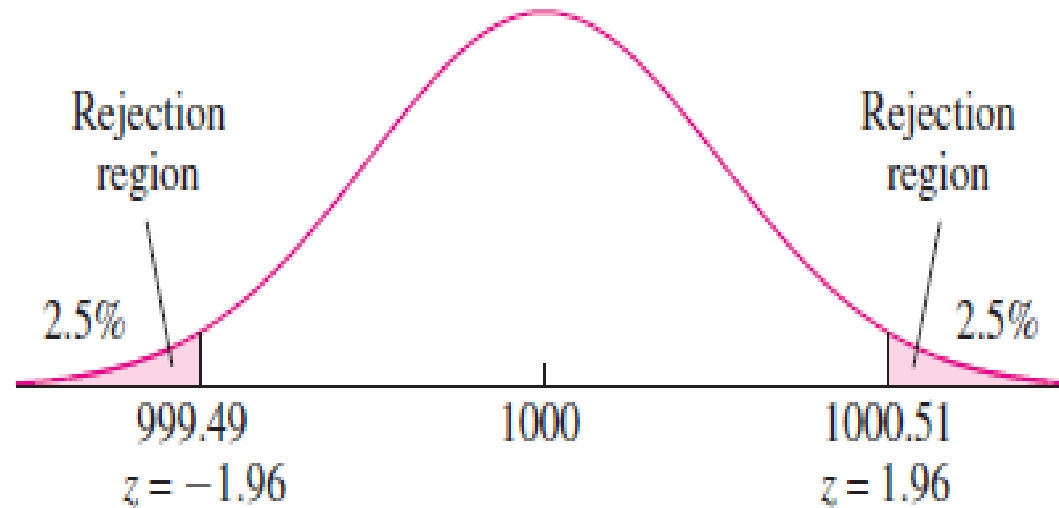


FIGURE 6.24 The rejection region for this two-tailed test consists of both the lower and the upper 2.5% of the null distribution. There are two critical points, 999.49 and 1000.51.

- The z-scores that cut off the upper and lower 2.5% of the distribution are $z = \pm 1.96$.
- Therefore the rejection region consists of all values of X
- greater than or equal to $1000 + (1.96)(0.26) = 1000.51$,
- along with all the values less than or equal to $1000 - (1.96)(0.26) = 999.49$.
- Note that there are two critical points, 999.49 and 1000.51.

Type I and Type II Errors

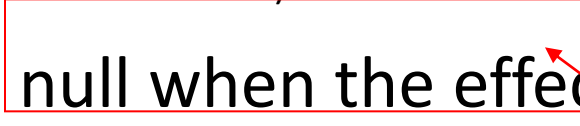
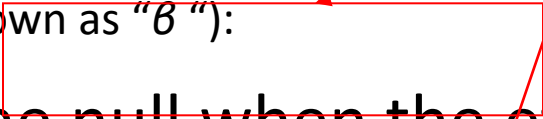

Summary

When conducting a fixed-level test at significance level α , there are two types of errors that can be made. These are

- Type I error: Reject H_0 when it is true.
- Type II error: Fail to reject H_0 when it is false.

The probability of a type I error is never greater than α .

Error and Power

- **Type-I Error** (also known as " α "): 
 - Rejecting the null when the effect isn't real.
- **Type-II Error** (also known as " β "): 
 - Failing to reject the null when the effect is real.
- **POWER** (the flip side of type-II error: $1 - \beta$): 
 - The probability of seeing a true effect if one exists.

Note the sneaky
conditionals...

Type I and Type II Error in a box

Your Statistical Decision	True state of null hypothesis	
	H_0 True (example: the drug doesn't work)	H_0 False (example: the drug works)
Reject H_0 (ex: you conclude that the drug works)	<i>Type I error</i> (α)	<i>Correct</i>
Do not reject H_0 (ex: you conclude that there is insufficient evidence that the drug works)	<i>Correct</i>	<i>Type II Error (β)</i>

Error and Power

- Type I error rate (or significance level): the probability of finding an effect that isn't real (false positive).
 - If we require $p\text{-value} < .05$ for statistical significance, this means that 1/20 times we will find a positive result just by chance.
- Type II error rate: the probability of missing an effect (false negative).
- Statistical power: the probability of finding an effect if it is there (the probability of not making a type II error).
 - When we design studies, we typically aim for a power of 80% (allowing a false negative rate, or type II error rate, of 20%).

If α is the significance level that has been chosen for the test, then the probability of a type I error is never greater than α .

- When designing experiments whose data will be analyzed with a fixed-level test, it is important to try to make the probabilities of type I and type II errors reasonably small.
- compute the probability of a type I error and show that it is no greater than 0.05

- Let X_1, \dots, X_n be a large random sample from a population with mean μ and variance σ^2 . Then \bar{X} is normally distributed with mean μ and variance σ^2/n .
- Assume that we are to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ at the fixed level $\alpha = 0.05$.
- That is, we will reject H_0 if $P \leq 0.05$.
- The null distribution, shown in Figure, is normal with mean 0 and
- var
- Assi



FIGURE 6.25 The null distribution with the rejection region for $H_0 : \mu \leq 0$.

- A type I error will occur if we reject H_0 , which will occur if $P \leq 0.05$, which in turn will occur if $X \geq 1.645\sigma_X$.
- Therefore the rejection region is the region $X \geq 1.645\sigma_X$.
- Now since H_0 is true, $\mu \leq 0$.
- First, we'll consider the case where $\mu = 0$.
- Then the distribution of X is given by Figure. In this case, $P(X \geq 1.645\sigma_X) = 0.05$, so the
- probability of rejecting H_0 and making a type I error is equal to 0.05.
- Next, consider the case where $\mu < 0$.
- Then the distribution of X is obtained by shifting the curve in Figure to the left, so $P(X \geq 1.645\sigma_X) < 0.05$, and the probability of a type I error is less than 0.05.
- We could repeat this illustration using any number α in place of 0.05.
- We conclude that if H_0 is true, the probability of a type I error is never greater than α .
- Furthermore, note that if μ is on the boundary of H_0 ($\mu = 0$ in this case), then the probability of a type I error is equal to α .

- the smaller we make the probability of a type I error, the larger the probability of a type II error becomes.
- The usual strategy is to begin by choosing a value for α so that the probability of a type I error will be reasonably small.
- a conventional choice for α is 0.05.
- If the probability of a type II error is large, it can be reduced only by redesigning the experiment—for example by increasing the sample size.
- Calculating and controlling the size of the type II error is somewhat more difficult than calculating and controlling the size of the type I error.

Statistical Power

- The probability of rejecting a false null hypothesis (H_0).
- The probability of obtaining a value of t (or z) that is large enough to reject H_0 when H_0 is actually false
- We always test the null hypothesis against an alternative/research hypothesis
- Usually the goal is to reject the null hypothesis in favor of the alternative

Why is Power Important?

- As researchers, we put a lot of effort into designing and conducting our research. This effort may be wasted if we do not have sufficient power in our studies to find the effect of interest.

Power

- ▶ power (π) + β = 1
- ▶ β = (1- π) = probability of accepting false H_0 (ie. reject true H_a)
 - probability of Type II error
 - false positive
- ▶ **Power (π)** = (1- β) = probability of detecting a difference when a difference does exist
 - probability of accepting true H_a (ie. reject false H_0)
 - how sensitive your test is to the existing difference between the compared samples

Statistical Power

- Power is the ability of a test to detect a real effect. It is measured as a probability that equals $1 - \beta$.

Researcher Decision	Actual State of Reality	
	H_0 is true	H_0 is false
Reject H_0	Type I error (α)	Correct Decision ($1 - \beta$)
Accept H_0	Correct Decision ($1 - \alpha$)	Type II error (β)

example of a power calculation

- Assume that a new chemical process has been developed that may increase the yield over that of the current process. The current process is known to have a mean yield of 80 and a standard deviation of 5, where the units are the percentage of a theoretical maximum. If the mean yield of the new process is shown to be greater than 80, the new process will be put into production.
- Let μ denote the mean yield of the new process. It is proposed to run the new process 50 times and then to test the hypothesis
- $H_0 : \mu \leq 80$ versus $H_1 : \mu > 80$ at a significance level of 5%.

- If H_0 is rejected, it will be concluded that $\mu > 80$, and the new process will be put into production. Let us assume that if the new process had a mean yield of 81, then it would be a substantial benefit to put this process into production. If it is in fact the case that $\mu = 81$, what is the power of the test, that is, the probability that H_0 will be rejected?

- Note : in order to compute the power, it is necessary to specify a particular value of μ , in this case $\mu = 81$, for the alternate hypothesis.
- The reason for this is that the power is different for different values of μ .
- if μ is close to H_0 , the power will be small, while if μ is far from H_0 , the power will be large.

Computing the power involves two steps:

- 1.** Compute the rejection region.
- 2.** Compute the probability that the test statistic falls in the rejection region if the alternate hypothesis is true. This is the power.

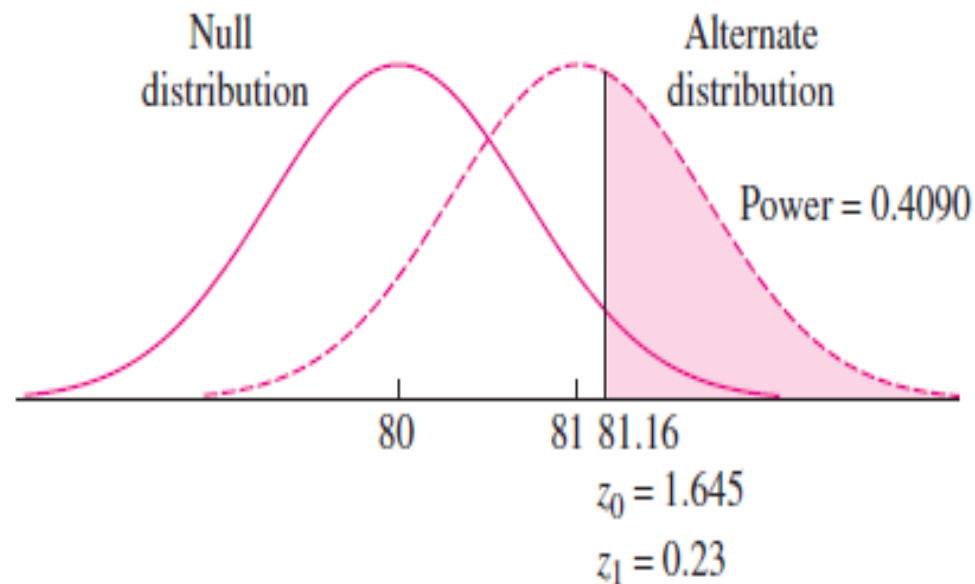


FIGURE 6.27 The rejection region, consisting of the upper 5% of the null distribution, is shaded. The z -score of the critical point is $z_0 = 1.645$ under the null distribution and $z_1 = 0.23$ under the alternate. The power is the area of the rejection region under the alternate distribution, which is 0.4090.

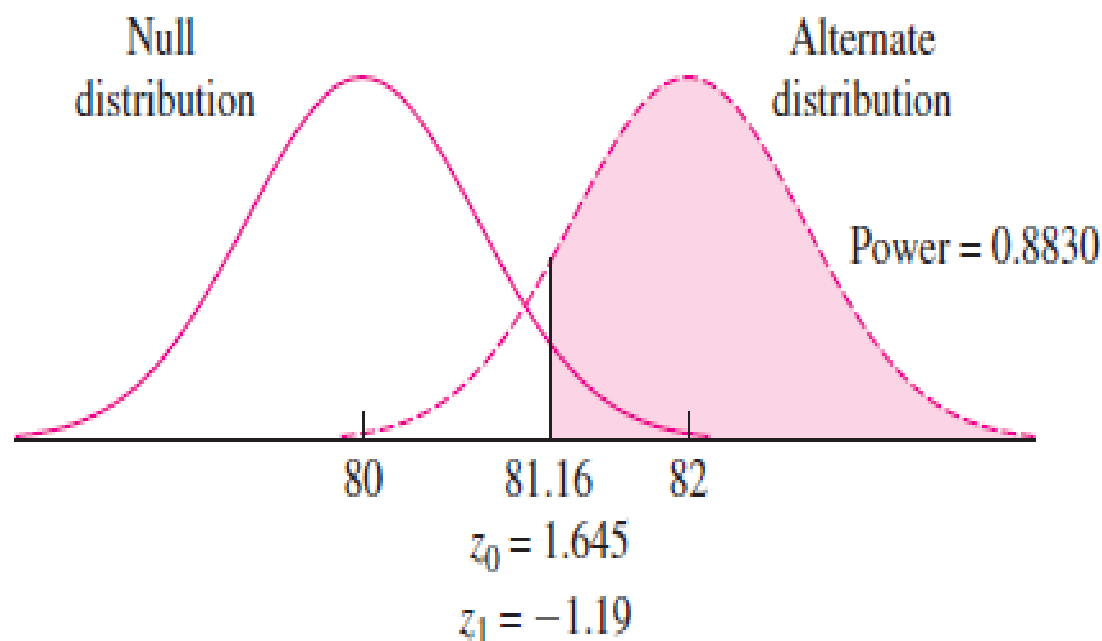


FIGURE 6.28 The rejection region, consisting of the upper 5% of the null distribution, is shaded. The z -score of the critical point is $z_0 = 1.645$ under the null distribution and $z_1 = -1.19$ under the alternate. The power is the area of the rejection region under the alternate distribution, which is 0.8830.

- Since the alternate distribution is obtained by shifting the null distribution, the power depends on which alternate value is chosen for μ , and can range from barely greater than the significance level α all the way up to 1.
- If the alternate mean is chosen very close to the null mean, the alternate curve will be almost identical with the null, and the power will be very close to α .
- If the alternate mean is far from the null, almost all the area under the alternate curve will lie in the rejection region, and the power will be close to 1.

- In testing the hypothesis $H_0 : \mu \leq 80$ versus $H_1 : \mu > 80$ regarding the mean yield of the new process, how many times must the new process be run so that a test conducted at a significance level of 5% will have power 0.90 against the alternative $\mu = 81$, if it is assumed that $\sigma = 5$?

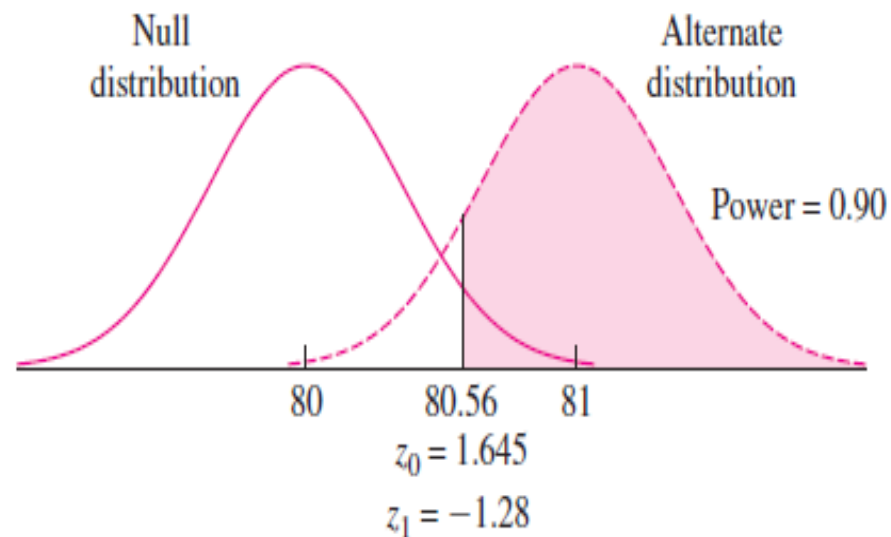


FIGURE 6.29 To achieve a power of 0.90 with a significance level of 0.05, the z -score for the critical point must be $z_0 = 1.645$ under the null distribution and $z_1 = -1.28$ under the alternate distribution.

Set them equal and solve for n .

$$80 + 1.645 \left(\frac{5}{\sqrt{n}} \right) = 81 - 1.28 \left(\frac{5}{\sqrt{n}} \right)$$

- Solving for n yields $n \approx 214$. The critical point can be computed by substituting this value for n into either side of the previous equation. The critical point is 80.56.

How do we know that power is large enough?

- Generally, the minimal sufficient (acceptable) value of power is 0.80



- $\pi \geq 0.80$

Analysis of power is performed:

- **1) Before gathering data**
- To determine the **minimal sample size** needed to have desired power in statistical testing (to detect a particular effect size)
- **2) After gathering data**
- To determine the **magnitude of power** that your statistical test will have given the sample parameters (**n** and **s**) and the magnitude of the effect that you want to detect

Power depends on:

- ▶ Sample size (n)
- ▶ Standard deviation (s)
- ▶ Alpha level (α)
- ▶ Size of effect/difference that you want to detect
- ▶ Type of statistical test performed

Power and Alpha (α)

- An increase in alpha, say from .05 to .1, artificially increases the power of a study.
- Increasing alpha reduces the risk of making a type II error, but increases that of a type I.
- Increasing the risk of making a type I error, in many cases, may be worse than making a type II error.
 - E.g., replacing an effective chemotherapy drug with one that is, in reality, less effective.

Power and Sample Size (N)

- Power increases as N increases.
- The more independent scores that are measured or collected, the more likely it is that the sample mean represents the true mean.
- Prior to a study, researchers rearrange the power calculation to determine how many scores (subjects or N) are needed to achieve a certain level of power (usually 80%).

Power and Effect Size

- Effect size is a measure of the difference between the means of two groups of data.
- For example, the difference in mean jump ht. between samples of vball and bball players.
- As effect size increases, so does power.
- For example, if the difference in mean jump ht. was very large, then it would be very likely that a Z or t-test on the two samples would detect that true difference.

A Little More on Effect Size

- While a p-value indicates the statistical significance of a test, the [effect size](#) indicates the “practical” significance.
- If the units of measurement are meaningful (e.g., jump height in cm), then the effect size can simply be portrayed as the difference between two means.
- If the units of measurement are not meaningful (questionnaire on behaviour), then a standardized method of calculating effect size is useful.