Hypothesis Testing

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:

1. W. Navidi. *Statistics for Engineering and Scientists*. Chapter 6 & Teaching Material

Introduction - Scenario for Hypothesis Test

- Recall the Example in Chapter 5 about Microdrills
  - Our sample had a mean $\overline{X}$ of 12.68 and standard deviation s of 6.83
  - Let us assume that the main question is: Whether or not the population mean lifetime μ is greater than 11 ?
    - We can address this by examining the value of the sample mean
    - We see that our sample mean is larger than 11, but because of uncertainty in the means, this does not guarantee that μ > 11
  - We would like to know just how certain we can be that μ > 11
    - A confidence interval is not quite what we need
  - The statement "μ > 11" is a hypothesis about the population mean μ
  - To determine just how certain we can be that a hypothesis is true, we must perform a hypothesis test (假設檢定)

Large-Sample Test for a Population Mean

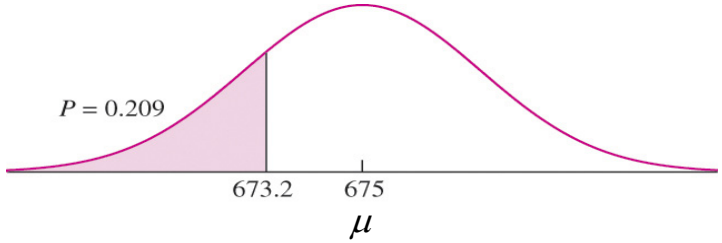
- The null hypothesis (虛擬假設)
  - (In most cases) Says that the effect indicated by the sample is due only to random variation between the sample and the population. We denote this with H_0
- The alternate hypothesis (對立假設)
  - Says that the effect indicated by the sample is real, in that it accurately represents the whole population. We denote this with H_1
- In performing a hypothesis test, we essentially put the null hypothesis on trial
  - We begin by assuming that H_0 is true just as we begin a trial by assuming a defendant to be innocent
  - The random sample provides the evidence
  - The hypothesis test involves measuring the strength of the disagreement between the sample and H_0 to produce a number between 0 and 1, called a *P*-value

P-Value

- The *P*-value measures the plausibility of H_0
- The smaller the *P*-value, the stronger the evidence is against H_0
- If the *P*-value is sufficiently small, we may be willing to abandon our assumption that H_0 is true and believe H_1 instead
- This is referred to as rejecting the null hypothesis

Steps in Performing a Hypothesis Test

1. Define H_0 and H_1
2. Assume H_0 to be true



3. Compute a test statistic
   - A test statistic is a statistic that is used to assess the strength of the evidence against H_0. A test that uses the *z*-score as a test statistic is called a *z*-test
4. Compute the *P*-value of the test statistic
   - The *P*-value is the probability, assuming H_0 to be true, that the test statistic would have a value whose disagreement with H_0 is as great as or greater than actually observed. The *P*-value is also called the observed significance level

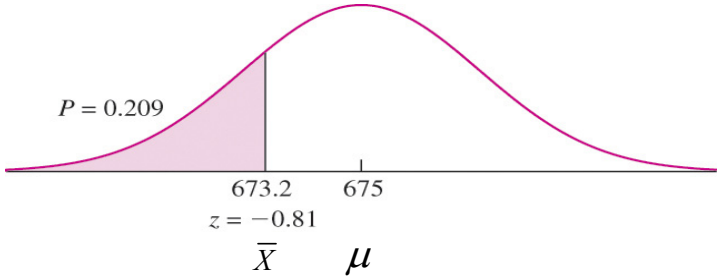
Example 6.1 (1/2)

- **Question**: A sample of 45 steel balls has average mean wear 673.2 μm and standard deviation 14.9 μm. Does the population of the steel balls have mean wear μ less than 675 μm ? Find the *P*-value for testing $H_0 : \mu \geq 675$ versus $H_1 : \mu < 675$
- **Answer**:
  - The null hypothesis ($\mu \geq 675$) is that μ is does not meet the specification. For this reason, values of the sample mean that are much smaller than μ will provide evidence against H_0
  - We assume that H_0 is true, and that therefore the sample readings were drawn from a population with mean $\mu = 675$ (the value closest to H_1). We approximate the population standard deviation with $s = 14.9$

Example 6.1 (2/2)

- The null distribution of $\overline{X}$ is normal with mean 1000 and standard deviation of $14.9/\sqrt{45} = 2.22$ The *z*-score of the observed $\overline{X} = 673.2$ is

$$z = \frac{673.2 - 675}{2.22} = -0.81$$



- Therefore if H_0 is true, there is a 20.9% chance to observe a sample whose disagreement with H_0 is as least as great as that was actually observed (i.e., $\overline{X}$)
- Since 0.209 is not a very small probability, we do not reject H_0
  - Instead, H_0 is plausible (note that we are not conclude that H_0 is true)

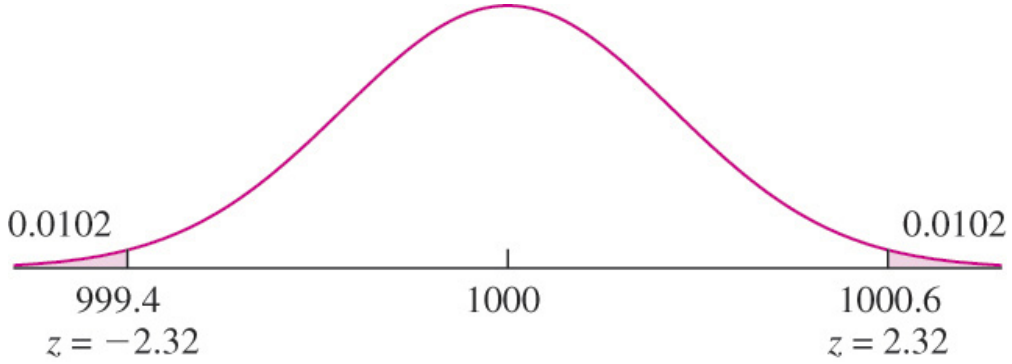
Example 6.2 (1/2)

- **Question**: A scale is to be calibrated by weighing a 1000 g test weight 60 times. The 60 scale readings have mean 1000.6 g and standard deviation 2 g. Find the *P*-value for testing $H_0 : \mu = 1000$ versus $H_1 : \mu \neq 1000$
- **Answer**:
  - The null hypothesis specifies that μ is equal to a specific value. For this reason, values of the sample mean that are either much larger or much smaller than μ will provide evidence against H_0
  - We assume that H_0 is true, and that therefore the sample readings were drawn from a population with mean $\mu = 1000$. We approximate the population standard deviation with $s = 2$

Example 6.2 (2/2)

- The null distribution of $\overline{X}$ is normal with mean 1000 and standard deviation of $2/\sqrt{60} = 0.258.$ The *z*-score of the observed $\overline{X} = 1000.6$ is

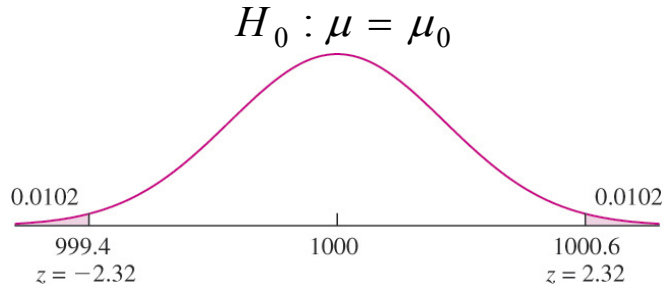
$$z = \frac{1000.6 - 1000}{0.258} = 2.32$$



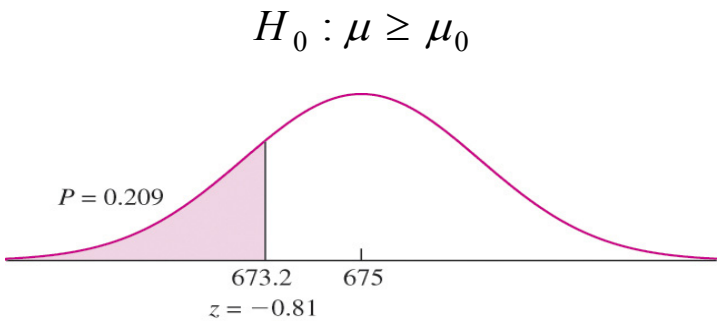
- Since H_0 specifies μ = 1000, regions in both tails of the curve are in greater disagreement with H_0 that the observed value of 1000.6. The *P*-value is the sum of the areas in both tails, which is 0.0204
- Therefore if H_0 is true, the probability of a result as extreme as or more extreme than that observed is only 0.0204
- The evidence against H_0 is pretty strong. It would be prudent to reject H_0 and to recalibrate the scale

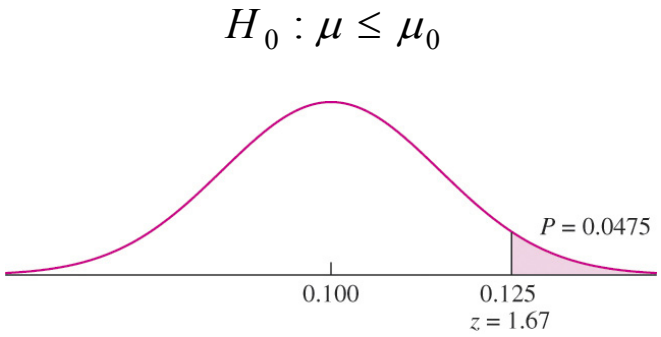
One and Two-Tailed Tests

- When H_0 specifies a single value for μ, both tails contribute to the *P*-value, and the test is said to be a two-sided or two-tailed test (雙尾檢定)



- When H_0 specifies only that μ is greater than or equal to, or less than or equal to a value, only one tail contributes to the *P*-value, and the test is called a one-sided or one-tailed test (單尾檢定)





Summary of *z*-test

- Let $X_1, \ldots, X_n$ be a *large* (e.g., $n > 30$) sample from a population with mean μ and standard deviation σ. To test a null hypothesis of the form H_0: $\mu \leq \mu_0$, H_0: $\mu \geq \mu_0$, or H_0: $\mu = \mu_0$
- Compute the *z*-score $z = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$
- If σ is unknown, it may be approximated by *s*
- Compute the *P*-value. The *P*-value is an area under the normal curve, which depends on the alternate hypothesis as follows:

  If the alternate hypothesis is H_1: $\mu > \mu_0$, then the *P*-value is the area to the right of *z*

  If the alternate hypothesis is H_1: $\mu < \mu_0$, then the *P*-value is the area to the left of *z*

  If the alternate hypothesis is H_1: $\mu \neq \mu_0$, then the *P*-value is the sum of the areas in the tails cut off by *z* and *-z*

Drawing Conclusions from the Results of Hypothesis Tests

- There are two conclusions that we draw when we are finished with a hypothesis test:
  - We reject H_0. In other words, we concluded that H_0 is false
  - We do not reject H_0. In other words, H_0 is plausible
- One can never conclude that H_0 is true. We can just conclude that H_0 might be true
- We need to know what level of disagreement, measured with the *P*-value, is great enough to render the null hypothesis implausible

More on the *P*-value

- The smaller the *P*-value, the more certain we can be that H_0 is false
- The larger the *P*-value, the more plausible H_0 becomes (but we can never be certain that H_0 is true)
- A rule of thumb suggests to reject H_0 whenever $P \leq 0.05$. While this rule is convenient, it has no scientific basis

Statistical Significance

- Whenever the *P*-value is less than a particular threshold, the result is said to be "statistically significant" at that level
  - So, if $P \leq 0.05$, the result is statistically significant at the 5% level
  - So, if $P \leq 0.01$, the result is statistically significant at the 1% level

- If the test is statistically significant at the 100α% level, we can also say that the null hypothesis is "rejected at level 100α%"

Example

- **Question**: A hypothesis test is performed of the null hypothesis H_0: $\mu = 0$. The *P*-value turns out to be 0.03. Is the result statistically significant at the 10% level? The 5% level? The 1% level? Is the null hypothesis rejected at the 10% level? The 5% level? The 1% level?

- **Answer**: The result is statistically significant at any level greater than or equal to 3%.
  - Thus it is statistically significant at the 10% and 5% level, but not at the 1% level
  - Similarly, we can reject the null hypothesis at any level greater than or equal to 3%, so H_0 is rejected at the 10% and 5% level, but not at the 1% level

Comments

- Some people report only that a test significant at a certain level, without giving the *P*-value. Such as, the result is "statistically significant at the 5% level," or "$P<0.05$." This is poor practice:
  - First, it provides no way to tell whether the *P*-value was just barely less than 0.05, or whether it was a lot less
  - Second, reporting that a result was statistically significant at the 5% level implies that there is a big difference between a *P*-value just under 0.05 and one just above 0.05, when in fact there is little difference
  - Third, a report like this does not allow readers to decide for themselves whether the *P*-value is small enough to reject the null hypothesis
- Reporting the *P*-value gives more information about the strength of the evidence against the null hypothesis allows each reader to decide for himself or herself whether to reject the null hypothesis

Significance

- When a result has a small *P*-value, we say that it is "statistically significant"
- In common usage, the word *significant* means "important"
- It is therefore tempting to think that statistically significant results must always be important
- Sometimes statistically significant results do not have any scientific or practical importance

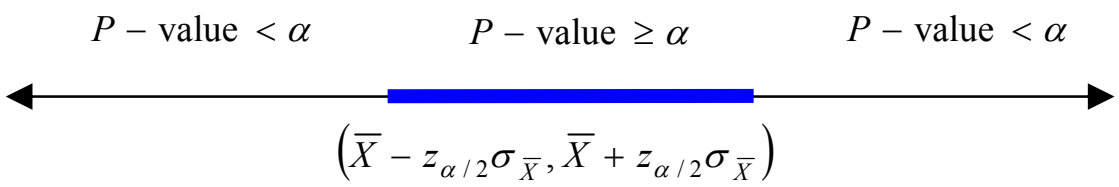
Hypothesis Tests and Confidence Intervals (1/2)

- Both confidence intervals and hypothesis tests are concerned with determining plausible values for a quantity such as a population mean μ
- In a hypothesis test for a population mean μ, we specify a particular value of μ (the null hypothesis) and determine that value is plausible
- A confidence interval for a population mean μ can be thought of as a collection of all values for μ that meet a certain criterion of plausibility, specified by the confidence level 100(1-α)%
- For example, the values contained within a two-sided level 100(1-α)% confidence intervals are precisely those values for which the *P*-value of a two-tailed hypothesis test will be greater than α

Hypothesis Tests and Confidence Intervals (2/3)

- Two-tailed Hypothesis Test

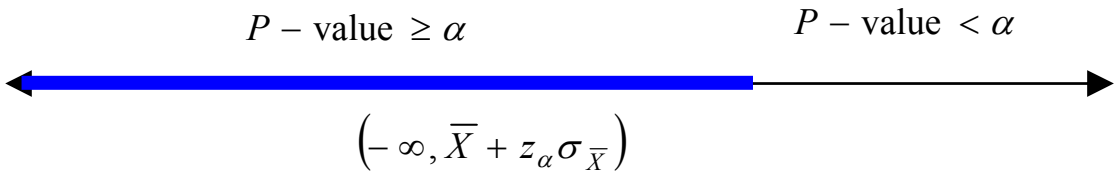
$$H_0 : \mu = \mu_0$$



$100(1-\alpha)\%$ confidence interval

- One-tailed Hypothesis Test (Low-tail, Upper Confidence Bound)

$$H_0 : \mu \geq \mu_0$$

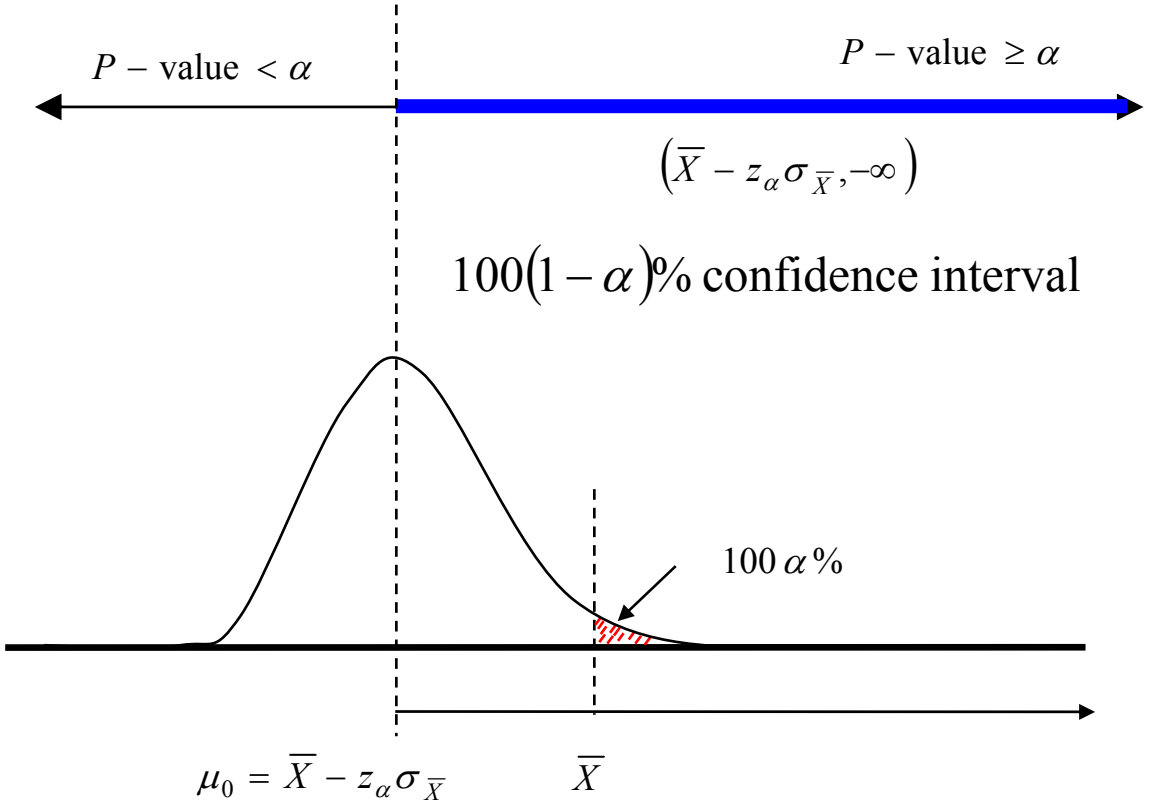


$100(1-\alpha)\%$ confidence interval

Hypothesis Tests and Confidence Intervals (2/3)

- One-tailed Hypothesis Test (Upper-tail, Low Confidence Bound)

$$H_0 : \mu \le \mu_0$$



Tests for a Population Proportion (1/2)

- A population proportion is simply a population mean for a population of 0's and 1's
  - A Bernoulli population
- We have a sample that consists of successes and failures
  - Thus, we have hypothesis concerned with a population proportion, it is natural to base the test on the sample proportion

Tests for a Population Proportion (2/2)

- Let X be the number of successes in n independent Bernoulli trials, each with success probability p; in other words, let $X \sim \text{Bin}(n, p)$
- To test a null hypothesis of the form $H_0: p \le p_0$, $H_0: p \ge p_0$, or $H_0: p = p_0$, assuming that both np_0 and $n(1- p_0)$ are greater than 10 (for normality):
  - Compute the z-score $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.
  - Compute the P-value. The P-value is an area under the normal curve, which depends on the alternate hypothesis as follows:
    - If the alternate hypothesis is $H_1: p > p_0$, the P-value is the area to the right of z
    - If the alternate hypothesis is $H_1: p < p_0$, the P-value is the area to the left of z
    - If the alternate hypothesis is $H_1: p \ne p_0$, the P-value is the sum of the areas in the tails cut off by z and $-z$

Small Sample Test for a Population Mean

- When we had a large sample we used the sample standard deviation *s* to approximate the population deviation σ
  - When the sample size is small, *s* may not be close to σ, which invalidates this large-sample method

- However, when the population is approximately normal, the Student's *t* distribution can be used
  - The only time that we don't use the Student's *t* distribution for this situation is when the population standard deviation σ is known. Then we are no longer approximating σ and we should use the *z*-test

Hypothesis Test

- Let $X_1, \ldots, X_n$ be a sample from a normal population with mean μ and standard deviation σ, where σ is unknown
- To test a null hypothesis of the form H_0: $\mu \leq \mu_0$, H_0: $\mu \geq \mu_0$, or H_0: $\mu = \mu_0$
  - Compute the test statistic $t_{n-1} = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$.
  - Compute the *P*-value The *P*-value is an area under the Student's *t* curve with $n - 1$ degrees of freedom, which depends on the alternate hypothesis as follows:
    - If the alternate hypothesis is H_1: $\mu > \mu_0$, then the *P*-value is the area to the right of *t*.
    - If the alternate hypothesis is H_1: $\mu < \mu_0$, then the *P*-value is the area to the left of *t*.
    - If the alternate hypothesis is H_1: $\mu \neq \mu_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and *-t*

Large Sample Tests for the Difference Between Two Means

- Now, we are interested in determining whether or not the means of two populations are equal

- The data will consist of two (independent) samples, one from each population

- We will compute the difference of the sample means
  - Since each of the sample means follows an approximate normal distribution, the difference is approximately normal as well
  - If the difference is far from zero, we will conclude that the population means are different

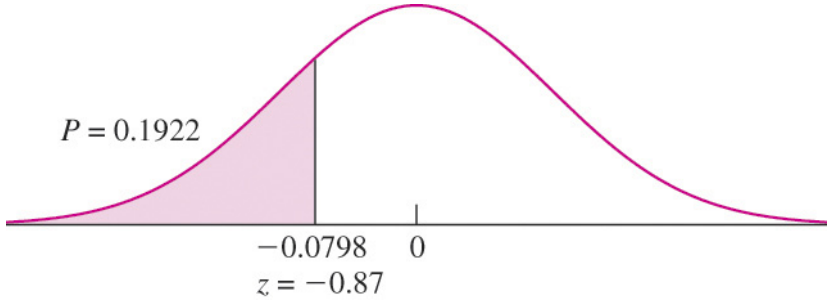
  - If the difference is close to zero, we will conclude that the population means might be the same

Hypothesis Test

- Let $X_1,\ldots,X_{n_X}$ and $Y_1,\ldots,Y_{n_Y}$ be *large* (e.g., $n_X > 30$ and $n_Y > 30$) samples from populations with mean μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Assume the samples are drawn independently of each other
- To test a null hypothesis of the form H_0: $\mu_X - \mu_Y \le \Delta_0$, H_0: $\mu_X - \mu_Y \ge \Delta_0$, or H_0: $\mu_X - \mu_Y = \Delta_0$
  - Compute the *z*-score $z = \dfrac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{\sigma_X^2 / n_X + \sigma_Y^2 / n_Y}}$.
    - If σ_X and σ_Y are unknown they may be approximated by s_X and s_Y.
  - Compute the *P*-value. The *P*-value is an area under the normal curve, which depends on the alternate hypothesis as follows:
    - If the alternate hypothesis is H_1: $\mu_X - \mu_Y > \Delta_0$, then the *P*-value is the area to the right of *z*
    - If the alternate hypothesis is H_1: $\mu_X - \mu_Y < \Delta_0$, then the *P*-value is the area to the left of *z*
    - If the alternate hypothesis is H_1: $\mu_X - \mu_Y \ne \Delta_0$, then the *P*-value is the sum of the areas in the tails cut off by *z* and *-z*

Tests for the Difference Between Two Proportions

- The procedure for testing the difference between two populations is similar to the procedure for testing the difference between two means
- We have random variables X and Y each with binomial distributions. $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$
- One of the null and alternative hypotheses are H_0: $p_X - p_Y \geq 0$ versus H_0: $p_X - p_Y < 0$



Comments

- The test is based on the statistic $\hat{p}_X - \hat{p}_Y$
  - We must determine the null distribution of this statistic
- By the central limit theorem, since n_X and n_Y are both large, we know that the sample proportions for *X* and *Y* have an approximately normal distribution
  - The difference between the proportions is also normally distributed
    - Whose mean is assumed to be equal to $p_X - p_Y = 0$
      - The population proportions are equal (i.e., p_X and p_Y should be estimated with a common value)
    - Estimate p_X and p_Y using the pooled proportion $\hat{p} = \frac{X+Y}{n_X + n_Y}$ then

$$\hat{p}_X - \hat{p}_Y \sim N\left(0, \hat{p}(1-\hat{p})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)\right)$$

Hypothesis Test

- Let $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$. Assume n_X and n_Y are large, and that X and Y are independent
- To test a null hypothesis of the form H_0: $p_X - p_Y \leq 0$, H_0: $p_X - p_Y \geq 0$, and H_0: $p_X - p_Y = 0$
  - Compute $\hat{p}_X = \frac{X}{n_X}, \hat{p}_Y = \frac{Y}{n_Y}, \text{and } \hat{p} = \frac{X+Y}{n_X + n_Y}.$
  - Compute the z-score $z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})(1/n_X + 1/n_Y)}}$ (assume mean of proportion difference is 0)
  - Compute the P-value. The P-value is an area under the normal curve, which depends on the alternative hypothesis as follows:
    - If the alternate hypothesis is H_1: $p_X - p_Y > 0$, then the P-value is the area to the right of z
    - If the alternate hypothesis is H_1: $p_X - p_Y < 0$, then the P-value is the area to the left of z
    - If the alternate hypothesis is H_1: $p_X - p_Y \neq 0$, then the P-value is the sum of the areas in the tails cut off by z and $-z$

Small Sample Tests for the Difference Between Two Means

- The *t* test can be used in some cases where samples are small, and thus the Central Limit Theorem does not apply

- If both populations are approximately normal, the Student's *t* distribution can be used to construct a hypothesis test

Hypothesis Test for Unequal Variance (1/2)

- Let $X_1,\ldots,X_{n_X}$ and $Y_1,\ldots,Y_{n_Y}$ be samples from *normal* populations with mean μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Assume the samples are drawn independently of each other
- Assume that σ_X and σ_Y are not known to be equal
- To test a null hypothesis of the form H_0: $\mu_X - \mu_Y \leq \Delta_0$, H_0: $\mu_X - \mu_Y \geq \Delta_0$, or H_0: $\mu_X - \mu_Y = \Delta_0$
  - Compute

$$v = \frac{\left[\left(s_X^2 / n_X\right) + \left(s_Y^2 / n_Y\right)\right]^2}{\left[\left(s_X^2 / n_X\right)^2 / (n_X - 1)\right] + \left[\left(s_Y^2 / n_Y\right)^2 / (n_Y - 1)\right]}$$

(degrees of freedom for *t* distribution)

rounded to the nearest integer

  - Compute the test statistic

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{s_X^2 / n_X + s_Y^2 / n_Y}}.$$

Hypothesis Test for Unequal Variance (2/2)

- Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with *v* degrees of freedom, which depends on the alternate hypothesis as follows:
  - If the alternate hypothesis is $H_1: \mu_X - \mu_Y > \Delta_0$, then the *P*-value is the area to the right of *t*
  - If the alternate hypothesis is $H_1: \mu_X - \mu_Y < \Delta_0$, then the *P*-value is the area to the left of *t*
  - If the alternate hypothesis is $H_1: \mu_X - \mu_Y \neq \Delta_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and *-t*

Hypothesis Test for Equal Variance (1/2)

- Let $X_1,\ldots,X_{n_X}$ and $Y_1,\ldots,Y_{n_Y}$ be samples from *normal* populations with mean μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively. Assume the samples are drawn independently of each other
- Assume that σ_X and σ_Y are known to be equal.
- To test a null hypothesis of the form H_0: $\mu_X - \mu_Y \le \Delta_0$, H_0: $\mu_X - \mu_Y \ge \Delta_0$, or H_0: $\mu_X - \mu_Y = \Delta_0$
  - Compute

$$s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}$$

  - Compute the test statistic (degrees of freedom for *t* distribution here are $\nu = n_X + n_Y - 2$)

$$t = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{s_p\sqrt{1/n_X + 1/n_Y}}.$$

Hypothesis Test for Equal Variance (2/2)

- Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with *v* degrees of freedom, which depends on the alternate hypothesis as follows:
  - If the alternate hypothesis is $H_1: \mu_X - \mu_Y > \Delta_0$, then the *P*-value is the area to the right of *t*
  - If the alternate hypothesis is $H_1: \mu_X - \mu_Y < \Delta_0$, then the *P*-value is the area to the left of *t*
  - If the alternate hypothesis is $H_1: \mu_X - \mu_Y \neq \Delta_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and *-t*

Tests with Paired Data

- Recall in in Chapter 5 that sometimes it is better to design a two-sample experiment so that each item in one sample is paired with an item in the other
- Here we present a method for testing hypotheses involving the difference between two population means on the basis of such paired data
- If the sample is large, the D_i (i.e., the population of differences) need not be normally distributed, the test statistic is

$$z = \frac{\overline{D} - \mu_0}{s_D / \sqrt{n}}$$

  - and a z-test should be performed

Hypothesis Test

- Let $(X_1, Y_1),\ldots, (X_n, Y_n)$ be sample of ordered pairs whose differences $D_1,\ldots,D_n$ are a sample from a *normal* population with mean μ_D
- To test a null hypothesis of the form H_0: $\mu_D \leq \mu_0$, H_0: $\mu_D \geq \mu_0$, or H_0: $\mu_D = \mu_0$
  - Compute the test statistic $t = \dfrac{\overline{D} - \mu_0}{s_D / \sqrt{n}}$
  - Compute the *P*-value. The *P*-value is an area under the Student's *t* curve with *n* – 1 degrees of freedom, which depends on the alternate hypothesis as follows
    - If the alternate hypothesis is H_1: $\mu_D > \mu_0$, then the *P*-value is the area to the right of *t*
    - If the alternate hypothesis is H_1: $\mu_D < \mu_0$, then the *P*-value is the area to the left of *t*
    - If the alternate hypothesis is H_1: $\mu_D \neq \mu_0$, then the *P*-value is the sum of the areas in the tails cut off by *t* and *-t*

Chi-Square (χ^2) Tests

- A generalization of the Bernoulli trial is the **multinomial trial**, which is an experiment that can result in any one of k outcomes, where $k \geq 2$.
  - We have already discussed this in Chapter 4

- Suppose we roll a six-sided die 600 times
  - The results obtained are called the **observed values**
  - To test the null hypothesis that $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$, we calculate the **expected values** for the given outcome
  - The idea behind the hypothesis test is that if H_0 is true, then the observed and expected values are likely to be close to each other

The Test

- Therefore we will construct a test statistic that measures the closeness of the observed to the expected values.
- The statistic is called the chi-square statistic
- Let k be the number of possible outcomes and let O_i and E_i be the observed and expected number of trials that result in outcome i
- The chi-square statistic is

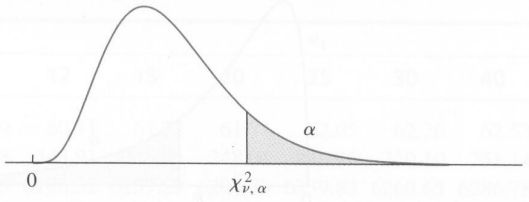
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Decision for a Test

- The larger the value of χ^2, the stronger the evidence against H_0
- To determine the *P*-value for the test, we must know the null distribution of this test statistic
- When the expected values are all sufficiently large, a good approximation is available. It is called the chi-square distribution with $k - 1$ degrees of freedom
  - Use of the chi-square distribution is appropriate whenever all the expected values are greater than or equal to 5
- A table for the chi-square distribution is provided in Appendix A, Table A.6

Chi-Square (χ^2) Distribution

TABLE A.6 Upper percentage points for the χ^2 distribution



	α									
ν	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.960	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766

For $\nu > 40$, $\chi^2_{\nu,\alpha} \approx 0.5(z_\alpha + \sqrt{2\nu - 1})$.

Example

- **Question**: A gambler rolls the 6-sided die 600 times to see whether the it deviates from fairness. Let the null hypothesis state that the die is fair, so the probabilities specified under null hypothesis is: $p_1 = p_2 = \cdots = p_6 = 1/6$

TABLE 6.3 Observed and expected values for 600 rolls of a die

Category	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Total	600	600

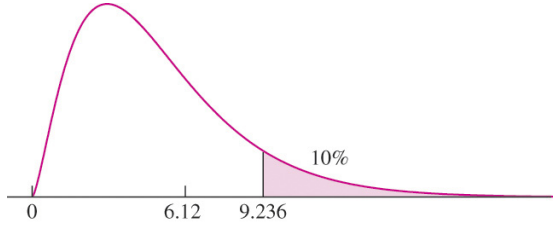
All the expected values ≥ 5, so the use of χ^2 distribution is appropriate

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\Rightarrow$$

$$\chi_5^2 = \frac{(115 - 100)^2}{100} + \cdots + \frac{(86 - 100)^2}{100}$$

$$= 6.12$$



The upper 10% point is 9.236, so we can conclude that P - value $> 10\%$.

$\therefore$ There is no evidence to suggest the die is not fair

Chi-Square (χ^2) Tests for Homogeneity (1/2)

- Sometimes several multinomial trials (experiments) are conducted, each with the same set of possible outcomes
- The null hypothesis is that the probabilities of the outcomes are the same for each experiment
- There is a chi-squared statistic for testing for homogeneity
- We can use a contingency table to record the observed values of different experiments
- The null hypothesis the proportion of different outcomes is the same for all experiments

Chi-Square (χ^2) Tests for Homogeneity (2/2)

- Given an contingency table shown below

TABLE 6.5 Notation for observed values

	Column 1	Column 2	⋯	Column *J*	Total
Row 1	O_{11}	O_{12}	⋯	O_{1J}	$O_{1.}$
Row 2	O_{21}	O_{22}	⋯	O_{2J}	$O_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
Row *I*	O_{I1}	O_{I2}	⋯	O_{IJ}	$O_{I.}$
Total	$O_{.1}$	$O_{.2}$	⋯	$O_{.J}$	$O_{..}$

I experiments

J outcomes

O_{ij} : observed value in cell ij

$O_{i.}$: the sum of observed values in row i

$O_{.j}$: the sum of observed values in column j

$O_{..}$: the sum of the observed values in all the cells

  - H_o: The null hypothesis the proportion of different outcomes is the same for all experiments ($i.e.$, for each column j, $p_{1j} = \cdots = p_{Ij}$)

proportion of outcome *j*

$$\Rightarrow p_{1j} = \cdots = p_{Ij} = \frac{O_{.j}}{O_{..}}$$

(the porportion)

$$E_{ij} = \frac{O_{.i}O_{.j}}{O_{..}}$$

(the expected number of trials whose outcomes falls into column j)

The test statistic is based on the differences between the observed and expected values

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

with $(I-1)(J-1)$ degree of freedom

Example 6.21

- **Question**: Four machines manufacture cylindrical steel pins. The pins are subject to a diameter specification.
- The null hypothesis says that the probabilities associated the pin categories are the same for the four machine.

TABLE 6.4 Observed numbers of pins in various categories with regard to a diameter specification

	Too Thin	OK	Too Thick	Total
Machine 1	10	102	8	120
Machine 2	34	161	5	200
Machine 3	12	79	9	100
Machine 4	10	60	10	80
Total	66	402	32	500

Expected values for Table 6.4

	Too Thin	OK	Too Thick	Total
Machine 1	15.84	96.48	7.68	120.00
Machine 2	26.40	160.80	12.80	200.00
Machine 3	13.20	80.40	6.40	100.00
Machine 4	10.56	64.32	5.12	80.00
Total	66.00	402.00	32.00	500.00

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$= 15.5844$$

$$\because P(\chi_6^2 > 14.449) = 2.5\% \text{ and } P(\chi_6^2 > 16.812) = 1\%$$

$$\therefore \quad 0.01 < P\text{-value} < 0.025$$

We can conclude that the machines differ in the proportions of pins that belong to different categories

Chi-Square (χ^2) Tests for Independence

- There is also a chi-square test for independence between rows and columns in a contingency table

- The null hypothesis is that the classification of the categories of one measurement do not depend on the classification of the categories of another measurement

Example 6.22

- **Question**: Test the dependence of the lengths and the diameters of cylindrical steel pins.
- The null hypothesis says that the diameter classification does not depend on the length classification (similar to the previous example in which we assume the diameter classification proportions are the same for all machines)

Observed Values for 1000 Steel Pins

Length	Diameter: Too Thin	OK	Too Thick	Total
Too Short	13	117	4	134
OK	62	664	80	806
Too Long	5	68	8	81
Total	80	849	92	1021

Expected Values for 1000 Steel Pins

Length	Diameter: Too Thin	OK	Too Thick	Total
Too Short	10.50	111.43	12.07	134.0
OK	63.15	670.22	72.63	806.0
Too Long	6.35	67.36	7.30	81.0
Total	80.0	849.0	92.0	1021.0

$$\chi_4^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= 7.46$$

$$\because P(\chi_4^2 > 7.779) = 10\%$$

$$\therefore P\text{-value} > 0.10$$

There is no evidence we can reject H_0

$\Rightarrow$ The length and the thickness (of the diameter) of the pins are independent!

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 hypothesis
- 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
  - It is best simply to report the *P*-value instead of making a firm decision whether or not to reject
- If a decision is going to be made on the basis of a hypothesis test, there is no choice but to pick a cut-off point for the *P*-value
  - When this is done, the test is referred to as a fixed-level test

Conducting Fixed-Level Testing

- 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

Comments on Fixed-Level Testing

- 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 more 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

Example 6.26

- **Question**: A sample of 100 concrete blocks are drawn. Assume the population standard deviation is σ close to 70MPa. If we have the hypothesis test that H_0: $\mu \leq 1350$ (MPa) and H_1: $\mu > 1350$ (MPa) (μ : population mean), find the rejection region when the test will be conducted at a significance level of 5% (one-tailed test)

The z - score of the point that cuts of 5% of the normal curve :

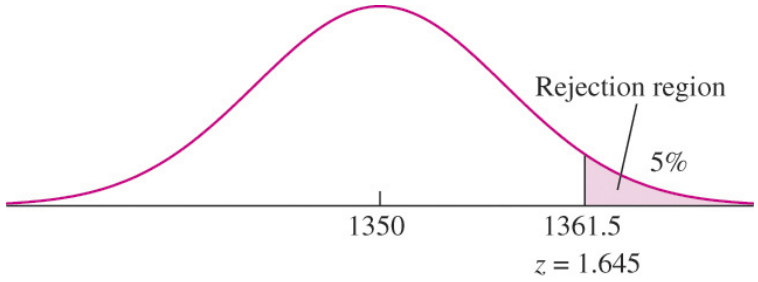
$$z_{0.05} = 1.645$$

The stand deviation of sample mean $\overline{X}$:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{70}{\sqrt{100}} = 7$$

Therefore, the critical point is

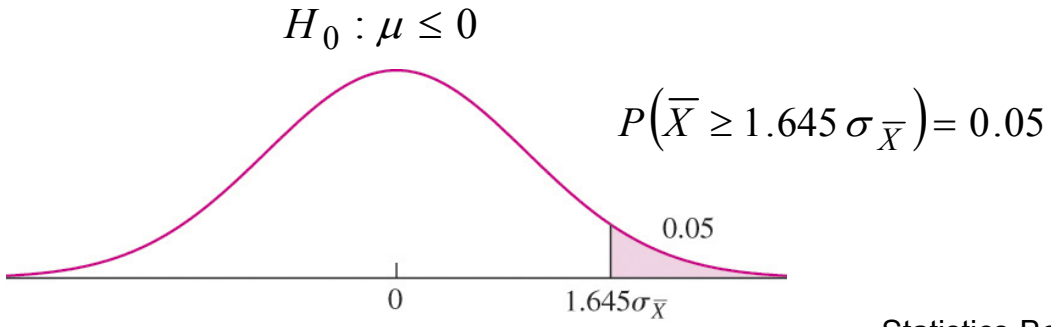
$$1350 + z_{0.05} \cdot \sigma_{\overline{X}}$$
$$= 1350 + 1.645 \cdot 7$$
$$= 1361.5$$



Errors of Fixed-Level Testing (1/2)

- 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 (False Rejection)
  - Type II error: Fail to reject H_0 when it is false (False Acceptance)

- We always try to make the probabilities of Type I and Type II errors reasonably small
- The probability of Type I error is never greater than α



Errors of Fixed-Level Testing (2/2)

- The smaller we make the probability of a type I error, the larger the probability of type II error becomes (why?)

Power

- A hypothesis test results in Type II error if H_0 is not rejected when it is false
- The power of the test is the probability of *rejecting* H_0 when it is false. Therefore,

  Power = 1 – P(Type II error)

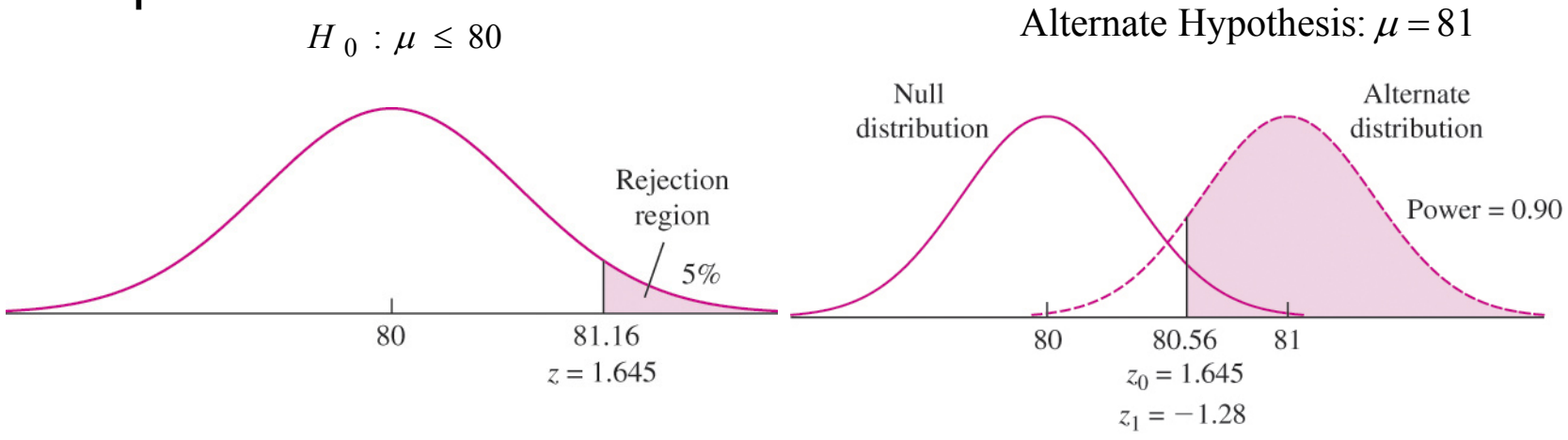
  - To be useful, a test must have reasonable small probabilities of both type I and type II errors
  - The type I error is kept small by choosing a small value of α as the significance level
  - If the power is large, then the probability of type II error is small as well, and the test is a useful one
- The purpose of a power calculation is to determine whether or not a hypothesis test, when performed, is likely to reject H_0 in the event that H_0 is false

Computing the Power

This involves two steps:

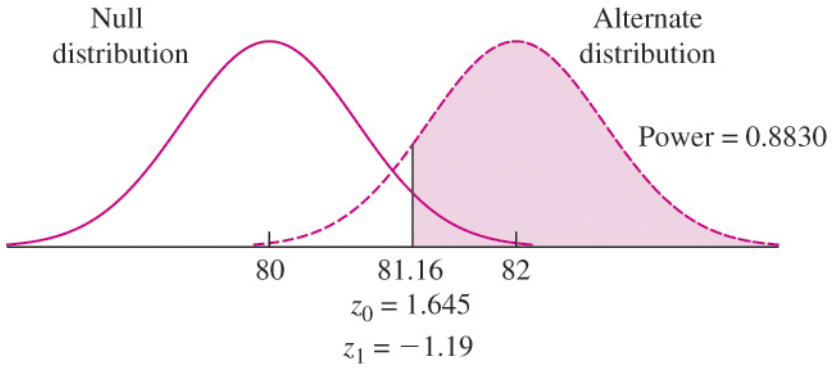
1. Compute the rejection region

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



Example 6.28

- **Question**: Find the power of 5% level test of H_0: $\mu \leq 80$ versus H_1: $\mu > 80$ for the mean yield of the new process under the alternative $\mu = 82$, assuming n (sample size)=50 and σ (population standard deviation)=5



We reject H_0 if the sample mean $\overline{X} \geq 81.16$

The z - score for the critical point of 81.16 under

The alternate hypothesis is $z = (81.16 - 82)/\left(5/\sqrt{50}\right) = -1.19$

The power is the area to the right of $z = -1.19$ (of the alternate hypothesis) which is 0.8830

Comments on Power

- The power depends on which alternate value is chosen, and can range from barely great than the significance level α (when the alternate is very close to the null) all the way up to 1 (when the alternate is far from the null)

- When power is not large enough, it can be increased by increasing the sample size

Summary

- We learned about:
  - Large sample tests for a population mean
  - Drawing conclusions from the results of hypothesis tests
  - Tests for a population proportion and differences in two proportions
  - Small sample tests for a population mean
  - Large and small sample tests for the difference between two means
  - Tests with paired data
  - Chi-Square test
  - Errors (type I & II) of hypothesis testing