Methods for Statistical Prediction
Financial Time Series I

Topic 1: Review on Hypothesis Testing

Hung Chen

Department of Mathematics

National Taiwan University

9/26/2002

OUTLINE

1. Fundamental Concepts
2. Neyman-Pearson Paradigm
3. Examples
4. Optimal Test
5. Observational Studies
6. Likelihood Ratio Test
7. One-sample and Two-sample Tests

Motivated Example on Hypothesis Testing

ESP experiment: guess the color of 52 cards with replacement.

- Experiment: Generate data to test the hypotheses.
- T: number of correct guess in 10 trials
- $H_0 : T \sim Bin(10, 0.5)$ versus $H_1 : T \sim Bin(10, p)$ with $p > 1/2$
- Consider the test statistic T and the rejection region $R = \{8, 9, 10\}$.
- Compute the probability of committing type 1 error:

$$\begin{aligned} \alpha &= P(R) = P(X > 7) \\ &= 0.0439 + 0.0098 + 0.0010 = 0.0547. \end{aligned}$$

- When rejection region= $R = \{7, 8, 9, 10\}$, $\alpha = P(X > 6) = 0.1172 + P(X > 7) = 0.1719$.
- Calculation of power when $R = \{8, 9, 10\}$. We compute what the power will be under various values of p.

$$p = 0.6 \quad P(X > 7 | p = 0.6) = 0.1673$$
$$p = 0.7 \quad P(X > 7 | p = 0.7) = 0.3828.$$

- Idea: A statistical test of a hypothesis is a rule which assigns each possible observation to one of two exclusive categories: *consistent with the hypothesis under consideration* and *not consistent with the hypothesis*.
- Will we make mistake?

Two Types of Error

		Reality	
		H_0 true	H_0 false
Test says	reject H_0	Type I Error	Good
	cannot reject H_0	Good	Type II Error

- Usually, P(Type I Error) is denoted by α and P(Type II Error) is denoted by β.
- In ESP experiment, α increases when R moves from $\{8, 9, 10\}$ to $\{7, 8, 9, 10\}$ but β decreases.

- Statistical Hypotheses testing is a formal means of choosing between two distributions on the basis of a particular statistic or random variable generated from one of them.
  - How do we accomodate the uncertainty on the observed data?
  - How do we evaluate a method?
- Neyman-Pearson Paradigm
  - Null hypothesis H_0
  - Alternate hypothesis H_A or H_1
  - The objective is to select one of the two based on the available data.
  - A crucial feature of hypothesis testing is that the two competing hypotheses are not treated in the same way: one is given the benefit of the doubt, the other has the burden of proof.
    The one that gets the benefit of the doubt is called the null hypothesis. The other is called the alternative hypothesis.
  - By definition, the default is H_0. When we carry out a test, we are asking whether

the available data is significant evidence in favor of H_1. We are not testing whether H_1 is true; rather, we are testing whether the evidence supporting H_1 is statistically significant.

– The conclusion of a hypothesis test is that we either reject the null hypothesis (and accept the alternative) or we fail to reject the null hypothesis.
  Failing to reject H_0 does not quite mean that the evidence supports H_0; rather, it means that the evidence does not strongly favor H_1.
  Again, H_0 gets the benefit of the doubt.

– Examples:

  * Suppose we want to determine if stocks picked by experts generally perform better than stocks picked by darts. We might conduct a hypothesis test to determine if the available data should persuade us that the experts do better. In this case, we would have
    H_0: experts not better than darts
    H_1: experts better than darts

* Suppose we are skeptical about the effectiveness of a new product in promoting dense hair growth. We might conduct a test to determine if the data shows that the new product stimulates hair growth. This suggests

  H_0: New product does not promote hair growth

  H_1: New product does promote hair growth

  Choosing the hypotheses this way puts the onus on the new product; unless there is strong evidence in favor of H_1, we stick with H_0.
* Suppose we are considering changing the packaging of a product in the hope of boosting sales. Switching to a new package is costly, so we will only undertake the switch if there is significant evidence that sales will increase. We might test-market the change in one or two cities and then evaluate the results using a hypothesis test. Since the burden of proof is on the new pack-

age, we should set the hypotheses as follows:

H_0: New package does not increase sales

H_1: New package does increase sales

– There are two types of hypotheses, simple ones where the hypothesis completely specifies the distribution.

– Simple hypotheses test one value of the parameter against another, the form of the distribution remaining fixed.

– Here is an example when they are both composite:

X_i: Poisson with unknown parameter

X_i is not Poisson

Steps for setting up test:

1. Define the null hypothesis H_0 (devil's advocate).
   Put the hypothesis that you don't believe as H_0
2. Define the alternative H_A (one sided /two sided).
3. Find the test statistic.
   Use heuristic or systematic methods.
4. Decide on the type I error: α that you are willing to take.
5. Compute the probability of observing the data given the null hypothesis: p-value.
6. Compare the p-value to α, if its smaller, reject H_0.

Example 1: Sex bias in graduate admission

- The graduate division of the University of California at Berkeley attempted to study the possibility that sex bias operated in graduate admissions in 1973 by examining admission data.
- In this case, what does the hypothesis of no sex bias corresponds to? It is natural to translate this into

$$P[Admit|Male] = P[Admit|Female].$$

- Data
  - There were $8,442$ men who applied for admission to graduate school that quarter, and $4,321$ women.
  - About 44% of the men and 35% of the women were admitted.
  - How do we perform this two-sample test?
- What is the conclusion?
  - two-sample test

$$\frac{0.44 - 0.35}{\sqrt{\frac{0.44 \times 0.56}{8442} + \frac{0.35 \times 0.65}{4321}}} = 9.948715.$$

– p-value is $1.283\exp(-22)$ when H_1 is a one-sided test $P[Admit|Male] > P[Admit|Female]$.

– p-value is $2.566\exp(-22)$ when H_1 is a two-sided test $P[Admit|Male] \neq P[Admit|Female]$.

Example 2: Effectiveness of Therapy

- Suppose that a new drug is being considered with a view to curing a certain disease.
- How do we evaluate its effectiveness?
- The drug is given to n patients suffering from the disease and the number x of cures is noted.
- We wish to test the hypothesis that there is at least a $50-50$ chance of a cure by this drug based on the following data:

$$x \text{ cures among } n \text{ patients.}$$

- Put the problem in the following framework of statistical test:
  - The sample space $\mathcal{X}$ is simple-it is the set $\{0, 1, 2, \ldots, n\}$. (i.e., X can take on $0, 1, 2, \ldots, n$.)
  - The family $\{P_\theta\}$ of possible distributions on $\mathcal{X}$ is (assuming independent patients) the family of binomial distributions, parametrized by the real parameter θ taking values in $[0, 1]$.

– θ is being interpreted as the probability of cure.

– $X \sim Bin(n, \theta)$

– The stated hypothesis defines the subset $\Theta_0 = [1/2, 1]$ of the parameter space.
$H_0 : \theta \geq 1/2$

– In this situation, only a small class of tests which seem worth considering on a purely intuitive basis.
We will only consider those for which the set of x taken to be consistent with Θ_0 have the form $\{x : x \geq k\}$

– **Question**: Does it make sense to consider that x cures out of n patients were consistent with Θ_0, while $x+1$ were not?

– What is a **reasonable** test?

A recipe: Optimal tests for simple hypotheses

- Null hypothesis $H_0 : f = f_0$
- Alternate hypothesis $H_A : f = f_1$
- Want to find a rejection region R such that the error of both types are as small as possible.

  $\int_R f_0(x)dx = \alpha \quad \text{and} \quad 1 - \beta = \int_R f_1(x)dx.$

- Neyman-Pearson Lemma:
  For testing $f_0(x)$ against $f_1(x)$ a critical region of the form

  $$\Lambda(x) = \frac{f_1(x)}{f_0(x)} \geq K$$

  where K is a constant has the greatest power (smallest β) in the class of tests with the same α.

  – Let R denote the rejection region determined by $\Lambda(x)$ and S denote the rejection of other testing procedure.
  – $\alpha_R = \int_R f_0(x)dx$, $\alpha_S = \int_S f_0(x)dx$,
  – $\alpha_R, \alpha_S \leq \alpha$

– $\beta_R - \beta_S = (\int_R - \int_S) f_1 dx = \int_{R \cap S^c} f_1 dx - \int_{S \cap R^c} f_1 dx$.
Since in R, $f_1 \geq f_0/K$ and in R_c, $-f_1 \geq -f_0/K$ we have:

$$\begin{aligned} \beta_R - \beta_S &\geq \frac{1}{K}\left(\int_{R \cap S^c} f_0 dx - \int_{S \cap R^c} f_0 dx\right) \\ &= \frac{1}{K}\left(\int_R f_0 dx - \int_S f_1 dx\right) = \frac{1}{K}(\alpha_R - \alpha_S) \end{aligned}$$

– When $\alpha_R = \alpha_S = \alpha$, $\beta_R - \beta_S \geq 0$.

Why Neyman-Pearson framework is being accepted?

- A test whose error probabilities are as small as possible is clearly desirable.
  However, we cannot choose the critical region in such a way that $\alpha(\theta)$ and $\beta(\theta)$ are simultaneously uniformly minimized.
  By taking the critical region as the empty set, we can make $\alpha(\theta) = 0$ and by taking the critical region as the sample space, we can make $\beta(\theta) = 0$. Hence a test which uniformly minimized both error-probability functions would require to have zero error probabilities, and usually no such test exists.

- The modification suggested by Neyman and Pearson is based on the fact that in most circumstances our attitudes to the hypotheses Θ_0 and $\Theta - \Theta_0$ are different- we are often asking if there is sufficient evidence to reject the hypothesis Θ_0.
  In terms of the two possible errors this may be translated into the statement that often the **Type I error is more serious than**

the Type II error.

- We should *control the probability of the Type I error* at some pre-assigned small value α, and then, subject to this control, look for a test which **uniformly minimizes the function describing the probabilities of Type II error**.
- Is this asymmetry on (H_0, H_1) reasonable? Can you come up an example with business application?
  - Suppose we use this testing technique in searching for regions of the genome that resemble other regions that are known to have significant biological activity.
  - One way of doing this is to align the known and unknown regions and compute statistics based on the number of matches.
  - To determine significant values of these statistics a (more complicated) version of the following is done.
    Thresholds (critical values) are set so that if the matches occur at random and the

probability of a match is 1/2, then the probability of exceeding the threshold (type I) error is smaller than α.

– No one really believes that H_0 is true and possible types of alternatives are vaguely known at best, but computation under H_0 is easy.

Now we use the following example to motivate Neyman-Pearson lemma. We start from the simplest possible situation, that where Θ has only two elements θ_0 and θ_1, say, and where $\Theta_0 = \{\theta_0\}$, $\Theta - \Theta_0 = \{\theta_1\}$. Note that a hypothesis which specifies a set in the parameter space containing only one element is called a *simple* hypothesis. Thus we are now considering testing a simple null-hypothesis against a simple alternative. In this case, the power function of any test reduces to a single number, and we examine the question of the existence of a most-powerful test of given significance level α.

Revisit the example that x cures out of n patients when $n = 5$. We wish to test

$$H_0 : p = 0.5 \quad \text{versus} \quad H_1 : p = 0.3.$$

- The probability distribution of X is

$X = x$	0	1	2	3	4	5
$p = 0.5$	0.031	0.156	0.313	0.313	0.156	0.031
$p = 0.3$	0.168	0.360	0.309	0.132	0.028	0.003
$f_1(x)/f_0(x)$	5.419	2.308	0.987	0.422	0.179	0.097

- Think of the meaning of likelihood ratio $f_1(x)/f_0(x)$.

- We consider all possible nonrandomized tests of significance level 0.2.

critical region	α	$1-\beta$	critical region	α	$1-\beta$
$\{0\}$	0.031	0.168	$\{0,1\}$	0.187	0.528
$\{1\}$	0.156	0.360	$\{0,4\}$	0.187	0.196
$\{4\}$	0.156	0.028	$\{1,5\}$	0.187	0.363
$\{5\}$	0.031	0.003	$\{4,5\}$	0.187	0.031
$\{0,5\}$	0.062	0.171			

- The best test is the one with critical region $\{0,1\}$. Can you give a reason for that? Or, can you find a rule?
  Try to think in terms of likelihood ratio by noting

$$f_1(x) = \frac{f_1(x)}{f_0(x)} \cdot f_0(x).$$

  As a hint, compare the two tests $\{0,1\}$ and $\{0,4\}$ with the same α. Observe that their

power are

$$\begin{aligned}\beta_{\{0,1\}} &= [P_{\{p=0.3\}}(r=0)] + P_{\{p=0.3\}}(r=1)\\ \beta_{\{0,4\}} &= [P_{\{p=0.3\}}(r=0)] + P_{\{p=0.3\}}(r=4).\end{aligned}$$

Compare $P_{\{p=0.3\}}(r=4)$ to $P_{\{p=0.3\}}(r=1)$.

- Conclusion: The critical region determined by $\{x : f_1(x)/f_0(x) \geq c\}$ is quite intuitive. Suppose that we set out to order points in the sample space according to the amount of evidence they provide for P_1 rather than P_0. We should naturally order them according to the value of the ratio $f_1(x)/f_0(x)$; any x for which this ratio is large provides evidence than P_1 rather than P_0 is the true underlying probability distribution. The Neyman-Pearson analysis gives us a basis for choosing c so that

$$P_1\left\{x : \frac{f_1(x)}{f_0(x)} \geq c\right\} = \alpha.$$

Now we use the Neyman-Pearson lemma to derive UMP test in the following two examples.

Example 3. Suppose that X is a sample of size 1. We wish to test whether it comes from

$N(0,1)$ or the double exponential distribution $DE(0,2)$ with the pdf $4^{-1}\exp(-|x|/2)$.

- Make a guess on the testing procedure?
- Since $P(f_1(x) = cf_0(x)) = 0$, there is a unique nonrandomized UMP test.
- The UMP test $T_*(x) = 1$ if and only if

$$\frac{\pi}{8}\exp(x^2 - |x|) > c^2$$

for some $c > 0$, which is equivalent to $|x| > t$ or $|x| < 1 - t$ for some $t > 1/2$.

- Suppose that $\alpha < 1/4$. We use

$$\alpha = E_0[T_*(X)] = P_0(|X| > t) = 0.3374 > \alpha.$$

Hence t should be greater than 1 and

$$\alpha = \Phi(-t) + 1 - \Phi(t).$$

Thus, $t = \Phi^{-1}(1-\alpha/2)$ and $T_*(X) = I_{(t,\infty)}(|X|)$.

- Why the UMP test rejects H_0 when $|X|$ is large?
- The power of T_* under H_1 is

$$E_1[T_*(X)] = P_1(|X| > t) = 1 - \frac{1}{4}\int_{-t}^{t} e^{-|x|/2}dx = e^{-t/2}.$$

Example 4. Let $X_1, \ldots, X_n$ be iid binary random variables with $p = P(X_1 = 1)$. Suppose that we wish to test $H_0 : p = p_0$ versus $H_1 : p = p_1$, where $0 < p_0 < p_1 < 1$.

- Since $P(f_1(x) = cf_0(x)) \neq 0$, we may need to consider randomized UMP test.
- A UMP test of size α is

$$T_*(Y) = \begin{cases} 1 & \lambda(Y) > c \\ \gamma & \lambda(Y) = c \\ 0 & \lambda(Y) < c, \end{cases}$$

  where $Y = \Sigma_{i=1}^n X_i$ and

$$\lambda(Y) = \left(\frac{p_1}{p_0}\right)^Y \left(\frac{1-p_1}{1-p_0}\right)^{n-Y}.$$

- Since $\lambda(Y)$ is increasing in Y, there is an integer $m > 0$ such that

$$T_*(Y) = \begin{cases} 1 & Y > m \\ \gamma & Y = m \\ 0 & Y < m, \end{cases}$$

  where m and γ satisfy

  $\alpha = E_0[T_*(Y)] = P_0(Y > m) + \gamma P_0(Y = m)$.

- Since Y has the binomial distribution $Bin(n, p)$, we can determine m and γ from

$$\alpha = \sum_{j=m+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} + \gamma \binom{n}{m} p_0^m (1-p_0)^{n-m}.$$

- Unless

$$\alpha = \sum_{j=m+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

for some integer m, the UMP test is a randomized test.

- Do you notice that the UMP test T_* does not depend on p_1?
  - Neyman-Pearson lemma tells us that we should put those x into rejection region according to its likelihood ratio until the level of test achieves α.
  - Think of two hypothesis testing problems: The first one is $H_0 : p = p_0$ versus $H_1 : p = p_1$ and the second one is $H_0 : p = p_0$ versus $H_1 : p = p_2$ where $p_1 > p_0$ and $p_2 > p_0$.
  - For the above two testing problems, both their likelihood ratios increase as y increases.

- T_* is in fact a UMP test for testing $H_0 : p = p_0$ versus $H_1 : p > p_0$.

- Suppose that there is a test T_* of size α such that for every $P_1 \in \mathcal{P}$, T_* is UMP for testing H_0 versus the hypothesis $P = P_1$.
  Then T_* is UMP for testing H_0 versus H_1.

Example: Suppose we have reason to believe that the true average monthly return on stocks selected by darts is 1.5%. We want to choose between $H_0 : \mu = 1.5$ versus $H_1 : \mu \neq 1.5$, where μ^- is the true mean monthly return.

- We need to select a significance level α. Let's pick $\alpha = 0.05$. This means that there is at most a 5% chance that we will mistakenly reject H_0 if in fact H_0 is true (Type I error). It says nothing about the chances that we will mistakenly stick with H_0 if in fact H_1 is true (Type II error).

- Large sample hypothesis test. Let's suppose we have samples $X_1, \cdots, X_n$ with $n > 30$.

  - The first step in choosing between our hypotheses is computing the following test

statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

- If the null hypothesis is true, then the test statistic Z has approximately a standard normal distribution by the central limit theorem.
- The test: If $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$, we reject the null hypothesis; otherwise, we stick with the null hypothesis. (Recall that

- T-test for normal population.
  Suppose now that we don't necessarily have a large sample but we do have a normal population. Consider the same hypotheses as before.
  - Now our test statistic becomes

$$t == \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

  - Under H_0, the test statistic t has a t-distribution with $n - 1$
  - Consider the mean return on darts. Suppose we have $n = 20$ observations (the 1-

month contests) with a sample mean of -1.0 and a sample standard deviation of 7.2.

Our test statistic is -1.55. The threshhold for rejection is $t_{19,0.025} = 2.093$.

Example: Consider the effect of a packaging change on sales of a product. Let μ be the (unknown) mean increase in sales due to the change. We have data available from a test-marketing study. We will not undertake the change unless there is strong evidence in favor of increased sales. We should therefore set up the test like this: $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$.

- Note that this is a one-sided test.
- This formulation implies that a large X (i.e., large increases in sales in a test market) will support H_1 (i.e., cause us to switch to the new package) but negative values of X (rejecting decreased sales) support H_0.
- The packaging example: Suppose that based on test-marketing in 36 stores we observe a sample mean increase in sales of 13.6 units per week with a sample standard deviation

of 42.
Is the observed increase significant at level α=0.05? To answer this, we compute the test statistic $Z = 1.80$.
Our cutoffffis $z_\alpha = 1.645$. Since $Z > z$, the increase is significant.

Observational Studies

- An observational study on sex bias in admissions to the Graduate Division at the University of California, Berkeley, was carried out in the fall quarter of 1973. Bickel, P., OConnell, J.W., and Hammel, E. (1975) Is there a sex bias in graduate admissions? *Science* **187**, 398-404.
  - There were $8,442$ men who applied for admission to graduate school that quarter, and $4,321$ women.
  - About 44% of the men and 35% of the women were admitted.
  - Assuming that the men and women were on the whole equally well qualified (and there is no evidence to the contrary), the difference in admission rates looks like a very strong piece of evidence to show that men and women are treated differently in the admission procedure.
- Admissions to graduate work are made separately for each major. By looking at each major separately, it should have been possi-

ble to identify the ones which discriminated against the women.

  - In Berkeley, there are over a hundred majors.
  - Look at the six largest majors had over five hundred applicants each. (They together accounted for over one third of the total number of applicants to the campus.)
  - In each major, the percentage of female applicants who were admitted is roughly equal to the percentage of male applicants.
  - The only exception is major A, which appears to discriminate against men: it admitted 82% of the women, and only 62% of the men.
  - When a;; six majors are taken together, they admitted 44% of the male applicants, and only 30% of the females-the difference is 14%,

- Admissions data in the six largest majors

	Men		*Women*	
Major	Number of applicants	Percent admitted	Number of applicants	Percent admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	59	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

- What is going on? An explanation:
  - The first two majors were easy to get into. Over 50% of the men applied to these two.
  - The other four majors were much harder to get into. Over 90% of the women applied to these four.
  - There was an effect due to the choice of major, confounded with the effect due to sex. When the choice of major is controlled for, as in the above Table, there is little difference in the admissions rates for men or women.
- An experiment is *controlled* when the in-

vestigators determine which subjects will be the controls and which will get the treatment-for instance, by tossing a coin.

- Statisticians distinguish carefully between controlled experiments and *observational studies*.
  - Studies of the effects of smoking are necessarily observational-nobody is going to smoke for ten years just to please a statistician.
  - Many problems can be studied only observationally and all observational studies have to deal with the problems of confounding.
  - For the admission example, it is wrong to campus-wide choice of major. We have to make comparisons for homogeneous subgroups.
  - This was not a controlled, randomized experiment, however; sex was not randomly assigned to the applicants.
- An alternative analysis: Compare the weighted average admission rates for men and women.

Consider

$$\frac{933}{4526} \times 62\% + \frac{585}{4526} \times 63\% + \frac{918}{4526} \times 37\% \\ + \frac{792}{4526} \times 33\% + \frac{584}{4526} \times 28\% + \frac{714}{4526} \times 6\%$$

and etc which lead to 39% versus 43%.

Hypothesis Testing By Likelihood Methods

Example Let $X_1, \ldots, X_n$ be iid with $X_1 \sim N(\mu, 1)$.

- Test $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_0 > 0$.
- Construct a test with $\alpha = 0.05$ and $\beta = 0.2005$.
- Reject H_0 if $\sqrt{n}\bar{X}_n > 1.645$.
- Note that

  $$\beta = P(\sqrt{n}\bar{X}_n \leq 1.645 | \mu = \mu_0) = \Phi(1.645 - \sqrt{n}\mu_0).$$

- If $n \to \infty$ and μ_0 is a fixed positive constant, $\beta \to 0$.
- To ensure $\beta = 0.2005$, it requires that

  $$1.645 - \sqrt{n}\mu_0 = -0.84$$

  or $\mu_0 = 2.485 n^{-1/2}$.
- Do you notice that μ_0 will change with n which is no longer a fixed alternative?

Test Statistics for A Simple Null Hypothesis

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0 \in R^s$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$.

Likelihood Ratio Test

- A *likelihood ratio* statistic,

$$\Lambda_n = \frac{L(\boldsymbol{\theta}^0; \mathbf{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x})}$$

  was introduced by Neyman and Pearson (1928).

- Λ_n takes values in the interval $[0, 1]$ and H_0 is to be rejected for sufficiently small values of Λ_n.

- The rationale behind LR tests is that when H_0 is true, Λ_n tends to be close to 1, whereas when H_1 is true, Λ_n tends to be close to 0,

- The test may be carried out in terms of the statistic

$$\lambda_n = -2 \log \Lambda_n.$$

- For finite n, the null distribution of λ_n will generally depend on n and on the form of pdf of X.

- LR tests are closely related to MLE's.

- Denote MLE by $\hat{\boldsymbol{\theta}}$. For asymptotic analysis, expanding λ_n at $\hat{\boldsymbol{\theta}}$ in a Taylor series, we get

$$\lambda_n \;=\; -2 \left\{ - \sum_{i=1}^{n} \log f(X_i, \hat{\boldsymbol{\theta}}) + \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}^0) \right\}$$

$$= 2\left\{\frac{1}{2}(\boldsymbol{\theta}^0 - \hat{\boldsymbol{\theta}})^T \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j\partial\theta_k} \log f(x;\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)(\boldsymbol{\theta}^0 -\right.$$

where $\hat{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$.

- Since $\boldsymbol{\theta}^*$ is consistent,

$$\lambda_n = n(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)^T \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j\partial\theta_k} L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)+o_P(1).$$

  By the asymptotic normality of $\hat{\boldsymbol{\theta}}$ and

$$-n^{-1}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta_j\partial\theta_k} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta}^0),$$

  λ_n has, under H_0, a limiting chi-squared distribution on s degrees of freedom.

Example Consider the testing problem H_0 : $\theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on iid $X_1, \ldots, X_n$ from the uniform distribution $U(0, \theta)$.

- $L(\theta_0; \mathbf{x}) = \theta_0^{-n} 1_{\{x_{(n)}<\theta_0\}}$
- $\hat{\theta} = x_{(n)}$ (MLE) and $\sup_{\theta\in\Theta} L(\theta; \mathbf{x}) = x_{(n)}^{-n} 1_{\{x_{(n)}<\theta\}}$
- We have

$$\Lambda_n = \begin{cases} (X_{(n)}/\theta^0)^n & X_{(n)} \leq \theta_0 \\ 0 & X_{(n)} > \theta_0 \end{cases}$$

- Reject H_0 if $X_{(n)} > \theta_0$ or $X_{(n)}/\theta_0 < c^{1/n}$.
- What is the asymptotic distribution of λ_n?
- What is $P(n \log(X_{(n)}/\theta^0) \leq c)$ where $c < 0$? It is not a χ^2 distribution. (Why???)

Example Consider the testing problem $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ based on iid $X_1, \ldots, X_n$ from the normal distribution $N(\mu_0, \sigma^2)$.

- $L(\theta^0; \mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left[-\Sigma_i (x_i - \mu_0)^2 / 2\sigma_0^2\right]$
- $\hat{\sigma}^2 = n^{-1} \Sigma_i (x_i - \mu_0)^2$ (MLE) and

$$\sup_{\theta \in \Theta} L(\theta; \mathbf{x}) = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2).$$

- We have

$$\Lambda_n = \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left(\frac{n}{2} - \frac{\Sigma_i (x_i - \mu_0)^2}{2\sigma_0^2}\right)$$

or under H_0

$$\lambda_n = -n \left\{ \ln\left(\frac{1}{n} \sum_{i=1}^{n} Z_i^2\right) - \left[1 - \left(\frac{1}{n} \sum_{i=1}^{n} Z_i^2\right)\right] \right\},$$

where $Z_1, \ldots, Z_n$ are iid $N(0, 1)$.

- Fact: Using CLT, we have

$$\frac{n^{-1} \Sigma_{i=1}^{n} Z_i^2 - 1}{\sqrt{2/n}} \xrightarrow{d} N(0, 1)$$

or

$$\frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1\right)^2 \xrightarrow{d} \chi_1^2.$$

- Note that $\ln u \approx -(1-u) - (1-u)^2/2$ when u is near 1 and $n^{-1}\,\Sigma_{i=1}^{n} Z_i^2 \to 1$ in probability by LLN.
- A common question to be asked in Taylor's series approximation is that how many terms we should consider. In this example, it refers to the use of approximation $\ln u \approx -(1-u)$ as a contrast to the second order approximation we use. If we do use the first order approximation, we will end up the difficulty of finding $\lim_n a_n b_n$ when $\lim_n a_n = \infty$ and $\lim_n b_n = 0$.
- We conclude that λ_n has a limiting chi-squared distribution with 1 degree of freedom.

Hypothesis Test on a Population Mean

1. We begin with the simplest case of a test. Suppose we are inclined to believe that some (unknown) population mean μ has the value μ_0, where μ_0 is some (known) number. We have samples $X_1, \ldots, X_n$ from the underlying population and we want to test our hypothesis that $\mu = \mu_0$. Thus, we have

$$\begin{aligned} H_0 : &\quad \mu = \mu_0 \\ H_1 : &\quad \mu \neq \mu_0 \end{aligned}$$

   What sort of evidence would lead us to reject H_0 in favor of H_1? Naturally, a sample mean far from μ_0 would support H_1 while one close to μ_0 would not. Hypothesis testing makes this intuition precise.

2. This is a **two-sided** or **two-tailed** test because sample means that are very large **or** very small count as evidence against H_0. In a **one-sided** test, only values in one direction are evidence against the null hypothesis. We treat that case later.

3. Example: Suppose we have reason to believe that the true average monthly return on stocks selected by darts is 1.5%. (See *Dart Investment Fund* in the casebook for background and data.) We want to choose between

$$\begin{aligned} H_0 : &\quad \mu = 1.5 \\ H_1 : &\quad \mu \neq 1.5, \end{aligned}$$

   where μ is the true mean monthly return.

4. We need to select a significance level α. Let's pick $\alpha = .05$. This means that there is at most a 5% chance that we will mistakenly reject H_0 if in fact H_0 is true (Type I error). It says nothing about the chances that we will mistakenly stick with H_0 if in fact H_1 is true (Type II error).

5. **Large sample hypothesis test**. Let's suppose we have samples $X_1, \ldots, X_n$ with $n > 30$. The first step in choosing between our hypotheses is computing the following **test statistic:**

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

   I am temporarily assuming that we know σ.

6. Remember that we know μ_0 (it's part of the null hypothesis we've formulated), even though we don't know μ.

7. If the null hypothesis is true, then the test statistic Z has approximately a standard normal distribution.

8. Now we carry out the test: If $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$ we **reject** the null hypothesis; otherwise, we stick with the null hypothesis. (Recall that $z_{\alpha/2}$ is defined by the requirement that the area to the right of $z_{\alpha/2}$ under $N(0,1)$ is $\alpha/2$. Thus, with $\alpha = .05$, the cutoff is 1.96.)

9. Another way to express this is to say that we reject if $|Z| > z_{\alpha/2}$; i.e., we reject if the test statistic Z lands in the set of points having absolute value greather than $z_{\alpha/2}$. This set is called the **rejection region** for the test.

10. Every hypothesis test has this general form: we compute a test statistic from data, then check if the test statistic lands inside or outside the rejection region. The rejection region depends on α but not on the data.

11. Notice that saying
$$-z_{\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$$
is equivalent to saying
$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$
So, here is another way to think of the test we just did. We found a confidence interval for the mean μ and checked to see if μ_0 lands in that interval. If μ_0 lands inside, we don't reject H_0; if μ_0 lands outside, we do reject H_0.

12. This supports our intuition that we should reject H_0 if $\overline{X}$ is far from μ_0.

13. As usual, if we don't know σ we replace it with the sample standard deviation s.

14. **T-test for normal population**. Suppose now that we don't necessarily have a large sample but we do have a normal population. Consider the same hypotheses as before. Now our **test statistic** becomes
$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$

15. Under the null hypothesis, the test statistic t has a t-distribution with $n-1$ degrees of freedom.

16. Now we carry out the test. Reject if
$$t < -t_{n-1,\alpha/2} \quad \text{or} \quad t > t_{n-1,\alpha/2};$$
otherwise, do not reject.

17. As before, rejecting based on this rule is equivalent to rejecting whenever μ_0 falls outside the confidence interval for μ.

18. Example: Let's continue with the hypothesis test for the mean return on darts. As above, $\mu_0 = 1.5$ and $\alpha = .05$. Suppose we have $n = 20$ observations (the 1-month contests) with a sample mean of -1.0 and a sample standard deviation of 7.2. Our test statistic is therefore
$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{-1.0 - 1.5}{7.2/\sqrt{20}} = -1.55$$
The threshhold for rejection is $t_{19,.025} = 2.093$. Since our test statistic t has an absolute value smaller than the cutoff, we cannot reject the null hypothesis. In other words, based on a significance level of .05 the evidence does not significantly support the view that $\mu \neq 1.5$.

19. Keep in mind that the deck is stacked in favor of H_0; unless the evidence is very compelling, we stick with the null hypothesis. The smaller we make α, the harder it is to reject H_0.

20. If a test leads us to reject H_0, we say that **the results are significant at level α.**

P-Values

1. There is something rather arbitrary about the choice of α. Why should we use $\alpha = .05$ rather than .01, .10 or some other value? What if we would have rejected H_0 at $\alpha = .10$ but fail to reject it because we chose $\alpha = .05$? Should we change our choice of α?

2. Changing α after a hypothesis test is "cheating" in a precise sense. Recall that, by the definition of α, the probability of a Type I error is at most α. Thus, fixing α gives us a guarantee on the effectiveness of the test. If we change α, we lose this guarantee.

3. Nevertheless, there is an acceptable way to report what would have happened had we chosen a different significance level. This is based on something called the **p-value** of a test.

4. The **p-value** is the smallest significance level (i.e., the smallest α) at which H_0 would be rejected, for a given test statistic. It is therefore a measure of how significant the evidence in favor of H_1 is: the smaller the p-value, the more compelling the evidence.

5. Example: Consider the test of mean returns on stocks picked by darts, as above. To simplify the present discussion, let's suppose we have 30 data points, rather than 20. (See *Dart Investment Fund* in the casebook for background and data.) As before the hypotheses are

$$\begin{aligned} H_0: &\quad \mu = 1.5 \\ H_1: &\quad \mu \neq 1.5 \end{aligned}$$

Let's suppose that our sample mean $\overline{X}$ (based on 30 observations) is -0.8 and the sample standard deviation is 6.1. Since we are assuming a large sample, our test statistic is

$$Z = \frac{\overline{X} - 1.5}{s/\sqrt{n}} = \frac{-0.8 - 1.5}{6.1/\sqrt{30}} = -2.06$$

With a significance level of $\alpha = .05$, we get $z_{\alpha/2} = 1.96$, and our rejection region would be

$$Z < -1.96 \text{ or } Z > 1.96.$$

So, in this case, $Z = -2.06$ would be significant: it is sufficiently far from zero to cause us to reject H_0.

We now ask, what is the smallest α at which we would reject H_0, based on $Z = -2.06$. We are asking for the smallest α such that $-2.06 < -z_{\alpha/2}$; i.e., the smallest α such that $z_{\alpha/2} < 2.06$. To find this value, we look up 2.06 in the normal table. This gives us .9803. Now we subtract this from 1 to get the area to the right of 2.06. This gives us

$1 - .9803 = .0197$. This is the the smallest $\alpha/2$ at which we would reject H_0. To find the smallest α, we double this to get .0394. This is the p-value for the test. It tells us that we would have rejected the null hypothesis using any α greater than .0394.

6. To verify that this is correct, let's work backwards. Suppose we had chosen $\alpha = .0394$ in the first place. Our rejection cut-off would then have been $z_{\alpha/2} = z_{.0197}$. To find this value, we look up $1 - .0197 = .9803$ in the body of the normal table; we find that $z_{.0197} = 2.06$. So, we reject if $Z < -2.06$ or $Z > 2.06$. Since our test statistic was $Z = 2.06$, we conclude that the p-value .0394 is indeed the significance level at which our test statistic just equals the cutoff.

7. **Summary of steps to find p-value** in a two-sided, large-sample hypothesis test on a population mean:

   (a) Compute the test statistic Z.

   (b) Look up the value of $|Z|$ in the normal table.

   (c) Subtract the number in the table from 1.

   (d) Multiply by 2; that's the p-value.

8. In principle, to find a p-value based on a t-test we would follow the same steps; however, our t-table does not give us all the information we have in the normal table, so we cannot get the p-value exactly. We can only choose from the α values available on the table or interpolate between them.

9. Here is another interpretation of the p-value: it is the probability of observing the results actually observed if the null hypothesis were true. Thus, a small p-value implies that it would be very unusual to observe the results actually observed if the null hypothesis were true. This leads us to reject the null hypothesis.

10. Most statistical packages (including spreadsheets) automatically report a p-value when they carry out a hypothesis test. You can then compare the p-value with your own personal α to determine whether or not to reject H_0.

One-Sided Tests on a Population Mean

1. In the setting considered so far, the null hypothesis is $\mu = \mu_0$, with μ_0 a fixed value. Very large values of $\overline{X}$ and very small values both count as evidence against H_0. We now consider cases (which are actually more common) in which only values in one direction support the alternative hypothesis. The general setting is this:

$$\begin{aligned} H_0 : &\quad \mu \leq \mu_0 \\ H_1 : &\quad \mu > \mu_0 \end{aligned}$$

2. Example: Consider the effect of a packaging change on sales of a product. Let μ be the (unknown) mean increase in sales due to the change. We have data available from a test-marketing study. We will not undertake the change unless there is strong evidence in favor of **increased** sales. We should therefore set up the test like this:

$$\begin{aligned} H_0: &\quad \mu \leq 0 \\ H_1: &\quad \mu > 0 \end{aligned}$$

This formulation implies that a large $\overline{X}$ (i.e., large increases in sales in a test market) will support H_1 (i.e., cause us to switch to the new package) but negative values of $\overline{X}$ (reflecting decreased sales) support H_0.

3. The mechanics of this test are quite similar to those of the two-sided test. The first thing we do is compute a test statistic. Assuming a large sample, we compute

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$

If our significance level is α, we find the corresponding value z_α (not $z_{\alpha/2}$!). We reject H_0 if

$$Z > z_\alpha.$$

4. Why are we now using z_α rather than $z_{\alpha/2}$? The short answer is that we are doing a one-sided rather than a two-sided test; so, we need all the area α in one tail rather than split between two tails.

5. A better answer is this: By definition, α is the maximum probability of a Type I error. Recall that Type I means rejecting H_0 when H_0 is true. If z is our cutoff, then a Type I error means observing $Z > z$ even though H_0 is true. So, we want

$$P(Z > z) = \alpha \text{ when } H_0 \text{ is true.}$$

Recall that when H_0 is true, Z has a standard normal distribution; the value of z that makes $P(Z > z) = \alpha$ is precisely z_α.

6. If you understand the explanation just given, then you have appreciated the fundamental principles of hypothesis testing. If not, just remember that in a two-sided test you use $z_{\alpha/2}$ and in a one-sided test use z_α.

7. We return to the packaging example: Suppose that based on test-marketing in 36 stores we observe a sample mean increase in sales of 13.6 units per week with a sample standard deviation of 42. Is the observed increase significant at level $\alpha = .05$? To answer this, we compute the test statistic

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = (13.6 - 0)/(42/6) = 1.80.$$

Our cutoff is $z_\alpha = z_{.05} = 1.645$. Since $Z > z_\alpha$, the increase is significant.

8. The t-distribution modification follows the usual pattern. Assuming a normal population, the test statistic
$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$
has a t-distribution with $n - 1$ degrees of freedom. We reject the null hypothesis if $t > t_{n-1,\alpha}$.

9. The definition of a p-value is the same as before: it is the smallest α at which the results would be rejected. The only difference comes from the fact that our cutoff is now z_α rather than $z_{\alpha/2}$.

10. **Finding p-value** in a **one-sided**, large-sample hypothesis test on a population mean:

    (a) Compute the test statistic Z.

    (b) Look up the value of Z in the normal table.

    (c) Subtract the number in the table from 1; that's the p-value.

    We no longer multiply by 2.

11. If we reverse the inequalities in the hypotheses to get
$$\begin{aligned} H_0 &: \quad \mu \geq \mu_0 \\ H_1 &: \quad \mu < \mu_0 \end{aligned}$$
the steps are exactly the same as before, except that now we reject H_0 if
$$Z < -z_\alpha \text{ or } t < -t_{n-1,\alpha};$$
in other words, when we reverse the inequalities, large negative test statistics support H_1.

12. To compute the p-value in this case, look up $-Z$ rather than Z. The other steps are unchanged.

Hypothesis Test on a Proportion

1. We now turn to a hypothesis test for a proportion, always assuming a large sample. The mechanics of this case are essentially the same as those for a mean.

2. The general two-sided test is
$$\begin{aligned} H_0 &: \quad p = p_0 \\ H_1 &: \quad p \neq p_0 \end{aligned}$$
and the general one-sided test is
$$\begin{aligned} H_0 &: \quad p \leq p_0 \\ H_1 &: \quad p > p_0 \end{aligned}$$
In all cases, p is an unknown population parameter while p_0 is a number we pick in formulating our hypotheses and is thus known.

3. Example: Let's test whether at least 50% of men who use Rogaine can be expected to show minimal to dense growth. This makes $p_0 = .50$ and p the true (unknown) proportion. Our test is

$$\begin{aligned} H_0 \quad & p \leq 0.5 \\ H_1 \quad & p > 0.5 \end{aligned}$$

The burden of proof is on Rogaine to show that the proportion is greater than one-half.

4. To test the hypotheses, we compute the test statistic

$$Z = \frac{\hat{p} - p_0}{\sigma_{p_0}},$$

where

$$\sigma_{p_0} = \sqrt{\frac{p_0(1-p_0)}{n}};$$

in other words, σ_{p_0} is what the standard error of $\hat{p}$ would be if p were actually p_0.

5. In an Upjohn study, 419 men out of 714 had at least minimal growth, so $\hat{p} = .59$. Since $p_0 = 0.5$, we have

$$\sigma_{p_0} = \sqrt{\frac{0.5(1-0.5)}{714}} = .0183$$

Thus, our test statistic is

$$Z = (.59 - .50)/.0183 = 4.81.$$

We now reject the null hypothesis if $Z > z_\alpha$. Clearly, 4.81 is larger than z_α for any reasonable choice of α, so the results are very significant.

6. The procedure for finding a p-value in this setting is exactly the same as in the test of a mean. In the example just carried out, 4.81 is off the normal table. Since our normal table goes up to 0.9998, we know that the p-value is less than 0.0002; in fact, we could report it as 0.000 since it is 0 to three decimal places. This means that the evidence is so overwhelming that there would be virtually no chance of observing the results in the study if the null hypothesis were true.

7. In the case of a two-sided test, we use $z_{\alpha/2}$ for a cut-off rather than z_α. If we reverse the inequalities in H_0 and H_1, the rejection condition becomes $Z < -z_\alpha$ rather than $Z > z_\alpha$. (You don't need to memorize this; just think about which direction supports H_1 and which supports H_0.)

Tests on Differences of Population Means

1. Hypothesis testing is frequently used to determine whether observed differences between two populations are significant:

   - Is the observed difference in performance between experts and darts significant?

- Is the mean midterm score among foreign students really greater than that among non-foreign students, or can the observed difference be attributed to chance?
- Is a new drug treatment more effective than an existing one?
- Do full-page ads significantly increase sales compared to half-page ads?

In each case, we are comparing two population means. Assuming we have samples from both populations, we can test hypotheses about the difference between the means.

2. We refer to one population using X quantities and the other using Y quantities; the two means are μ_X and μ_Y. The general two-sided test for the difference of two means has the form

$$\begin{aligned} H_0: &\quad \mu_X - \mu_Y = D_0 \\ H_1: &\quad \mu_X - \mu_Y \neq D_0 \end{aligned}$$

where D_0 is a specified (known) value we are testing. The general one-sided test is

$$\begin{aligned} H_0: &\quad \mu_X - \mu_Y \leq D_0 \\ H_1: &\quad \mu_X - \mu_Y > D_0 \end{aligned}$$

or else the same thing with the inequalities reversed.

3. We concentrate on the case $D_0 = 0$; that is, testing whether there is any difference between the two means. This is the most interesting and important case.

4. As with confidence intervals, here we distinguish a **matched pairs** setting and an **independent** setting. We begin with matched pairs.

5. We assume that we have samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from the two populations. Each X_i and Y_i are matched — they need not be independent. However, we do assume that (X_1, Y_1) are independent of (X_2, Y_2), etc. We compute the average difference

$$\overline{d} = \overline{X} - \overline{Y}$$

and the sample standard deviation of the differences

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[(X_i - Y_i) - (\overline{X} - \overline{Y})]^2}.$$

From these we get the test statistic

$$Z = \frac{\overline{d} - D_0}{s/\sqrt{n}}$$

or, in case of samples from normal populations,

$$t = \frac{\overline{d} - D_0}{s/\sqrt{n}}.$$

In the one-sided case, we compare the test statistic with z_α or $t_{n-1,\alpha}$ accordingly; in a two-sided test we use $\pm z_{\alpha/2}$ and $\pm t_{n-1,\alpha/2}$.

6. Example: Let's compare experts and darts. The burden is on the experts to prove they are better than darts, so we have

$$\begin{aligned} H_0: &\quad \mu_X - \mu_Y \leq 0 \\ H_1: &\quad \mu_X - \mu_Y > 0 \end{aligned}$$

the X's refer to expert returns, the Y's to dart returns. From the 20 1-month contests, we get $\overline{d} = 5.3$ and $s = 6.8$. Our test statistic is

$$t = \frac{5.3 - 0}{6.8/\sqrt{20}} = 3.48$$

Using $\alpha = 0.05$, our cutoff is $t_{19,.05} = 1.729$. Since $t > 1.729$ we reject the null hypothesis: the evidence in favor of the experts is significant.

7. Caveat: "Significant" here only refers to the outcome of this test. It does not imply that the conclusion itself is sound. For example, we may have reason to believe that the experts were unduly aided by the publication of their picks. This interferes with the sampling mechanism that generated our 20 data points. The test above is premised on the data coming from a random sample. The test says nothing about the quality of the data itself.

8. Now we consider the case of independent samples. We have samples $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ from two populations. We assume these are all independent of each other. Consider the test

$$\begin{aligned} H_0: &\quad \mu_X - \mu_Y \leq D_0 \\ H_1: &\quad \mu_X - \mu_Y > D_0 \end{aligned}$$

From the sample means $\overline{X}$ and $\overline{Y}$ and the sample standard deviations s_X and s_Y, we compute the test statistic

$$Z = \frac{\overline{X} - \overline{Y} - D_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}},$$

assuming that the sample sizes n_X and n_Y are large. We reject H_0 if $Z > z_\alpha$. In a two-sided test, we reject if $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$.

9. Calculating p-values, whether for matched pairs or independent samples works exactly the same way as before. Remember to multiply by 2 in the two-sided case.

10. The case of t-based comparisons is similar but a bit more complicated. If the variances of the two populations are assumed equal, we use a pooled estimate of the standard deviation and a t statistic with $n_X + n_Y - 2$ degrees of freedom. See §9.1 of LBS.

Testing a Difference of Proportions

1. We now test the difference between two unknown population proportions, p_X and p_Y.

(a) Is Rogaine more effective in promoting hair growth than a placebo?

(b) Based on a random sample of 100 shoppers, can we conclude that the market share of Colgate toothpaste differs from that of Crest?

(c) Is the proportion of women being promoted in an organization significantly smaller than the proportion of men promoted?

2. Let's formulate these comparisons in more detail:

   (a) With p_X the proportion for Rogaine and p_Y the proportion for the placebo, we want to test

   $$\begin{aligned} H_0 &: \quad p_X - p_Y \leq 0 \\ H_1 &: \quad p_X - p_Y > 0 \end{aligned}$$

   (b) With p_X and p_Y the proportion of Colgate and Crest buyers, we are testing

   $$\begin{aligned} H_0 &: \quad p_X - p_Y = 0 \\ H_1 &: \quad p_X - p_Y \neq 0 \end{aligned}$$

   In other words, we are testing if the market shares are the same or different.

   (c) With p_X the chances a woman is promoted and p_Y the chances a man is promoted, we are testing

   $$\begin{aligned} H_0 &: \quad p_X - p_Y \geq 0 \\ H_1 &: \quad p_X - p_Y < 0 \end{aligned}$$

   In this formulation, the burden of proof is on the claim of discrimination against women.

3. In each case above, the null hypothesis can equivalently be taken to state that $p_X = p_Y$. We compute a test statistic under the assumption that the null hypothesis is true. If the two proportions are equal, the following is a pooled estimate of the common proportion:

   $$\hat{p}_0 = \frac{n_X \hat{p}_X + n_Y \hat{p}_Y}{n_X + n_Y}.$$

   The corresponding estimate of the standard error is

   $$s_{\hat{p}_0} = \sqrt{\hat{p}_0(1 - \hat{p}_0)\left(\frac{n_X + n_Y}{n_X n_Y}\right)}.$$

   Our test statistic is

   $$Z = \frac{\hat{p}_X - \hat{p}_Y}{s_{\hat{p}_0}}.$$

   For the three tests described above, we have

   (a) Reject H_0 if $Z > z_\alpha$

   (b) Reject H_0 if $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

   (c) Reject H_0 if $Z < z_\alpha$