

Hypothesis Testing

Outline

1. Introduction
2. Hypothesis test under normal distribution
 1. Mean
 2. Variance
 3. Power analysis
 4. Sample size determination
3. Hypothesis test under binomial distribution
4. Hypothesis test under Poisson distribution

Example

- A current area of research interest is the familial aggregation of cardiovascular risk factors in general and lipid(脂质) levels in particular.
- Suppose the “average” cholesterol (胆固醇) level in children is 175 mg/dL. A group of men who have died from heart disease within the past year are identified, and the cholesterol levels of their offspring are measured.
- Two hypotheses are considered:
 - $H_0: \mu = 175$
 - $H_1: \mu > 175$Where μ is the average cholesterol level of these children.

One-sample test for the mean

- The t test for $H_0: \mu = \mu_0$ versus
 - $H_1: \mu = \mu_1 < \mu_0$
 - $H_2: \mu = \mu_1 \neq \mu_0$
 - $H_3: \mu = \mu_1 > \mu_0$
- Test statistic
 - t test: $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ (with σ unknown)
 - z test: $t = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ (with σ known)
- Reject the null hypothesis when
 - $H_1: t < t_{n-1, \alpha}$
 - $H_2: |t| > t_{n-1, \alpha/2}$
 - $H_3: t > t_{n-1, \alpha}$

P-values

- Notice that we rejected the a test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$, how about 0.001?
- The **P-value** is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than would be observed by chance alone. If the P-value is small, then either H_0 is true and we have observed a rare event or H_0 is false.
- Guidelines for Judging the Significance of a p -value
 - If $.01 \leq p < .05$, then the results are *significant*.
 - If $.001 \leq p < .01$, then the results are *highly significant*.
 - If $p < .001$, then the results are *very highly significant*.
 - If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by **NS**).

Criticisms of the P-value

- By reporting a P-value the reader can perform the hypothesis test at whatever level he or she chooses.
- However, P-values only consider significance
- It is difficult with a P-value or result of a hypothesis test to distinguish **practical significance** from **statistical significance**
 - i.e., the actual difference between \bar{x} and μ_0 may not be very large, but the results are statistically significant because of a large sample size.
 - See example in next slide

Example

- Suppose in the Obstetrics example given before, the mean birthweight was 120 oz. with a known sample standard deviation (S) of 24 oz. in the U.S.A. Assess the results of the study under the following two conditions.

- Suppose the mean birthweight was 119 oz., based on a sample size 10,000. The test statistic would be given by

$$t = \frac{119 - 120}{24/\sqrt{10,000}} = -4.17$$

very highly significant
but not important

Thus the p-value is given by $P(t_{9999} < -4.17) < .001$.

- Suppose the mean birthweight was 110 oz., based on a sample size of 10. The test statistic would be given by

$$t = \frac{110 - 120}{24/\sqrt{10}} = -1.32$$

not statistically
significant but could
be important

Thus the p-value is approximate by $P(t_9 < -1.32) = .110$

continued

- There is a close relation between confidence intervals and hypothesis testing:
 $p < 0.05$ (i.e. significant) \Leftrightarrow the 95% interval does not include the value specified in H_0 .
- The reason for this relation is that both methods are based on similar aspects of the theoretical distribution of the test statistic.
- The confidence interval shows the uncertainty, or lack of precision, in the estimate of interest, and thus conveys more useful information than the p-value.

continued

- The use of a new treatment is dependent not only on the significance but also on the amount of the effect. A single number (p-value) cannot convey the necessary information.
- P-values have become abusively used.
 - Unfortunately, some researchers have become polarized on this issue, with some statisticians favoring only the hypothesis-testing approach and some epidemiologists favoring only the CI approach.
- Don't just report P-values, give CIs too!
 - to provide complementary information

Two-sided versus one-sided(1)

- Test the hypothesis that the mean cholesterol level of recent female Asian immigrants is different from the mean in the general U.S. population.
 - $H_0: \mu = \mu_0 = 190$ vs. $H_1: \mu \neq \mu_0$
 - the two-sided p-value = $2 \times \Pr(t_{99} < -2.12) = 0.037$
- Suppose we guess from a previous review of the literature that the cholesterol level of Asian immigrants is likely to be **lower** than that of the general U.S. population because of better dietary habits.
 - use a one-sided test of the form $H_0: \mu = 190$ vs. $H_1: \mu < 190$.
 - the one-sided p-value = $\Pr(t_{99} < -2.12) = 0.018$
- Suppose we guess from a previous literature review that the cholesterol level of Asian immigrants is likely to be **higher** than that of the general U.S. population because of more stressful living conditions.
 - use a one-sided test of the form $H_0: \mu = 190$ vs. $H_1: \mu > 190$.
 - the p-value = $\Pr(t_{99} < -2.12) = .982$

Two-sided versus one-sided(2)

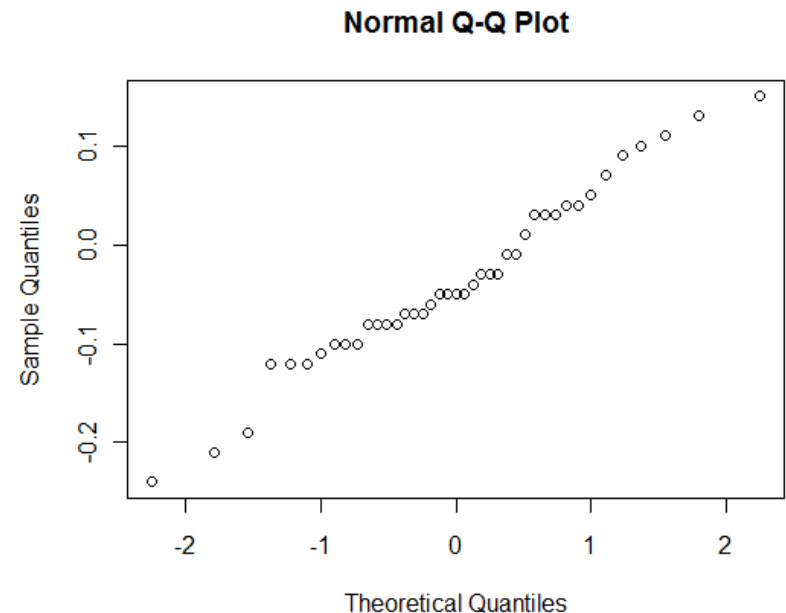


- One sided tests are rarely appropriate and in most cases two-sided tests are used. Even when there is strong prior expectations, for example that a new treatment can't be worse than the old one you can not be sure (otherwise you would not need an experiment).
- In all instances, it is important to decide whether to use a one-sided or a two-sided test **before data analysis** (or preferably before data collection) begins so as not to bias conclusions based on results of hypothesis testing. In particular, do not change from a two-sided to a one-sided test **after looking at the data**.

BMD Example

- The bone-mineral density (BMD) of the lumbar spine (ls) between heavier- and lighter-smoking twins is studied.
- Whether the data is normal (Or approximately normal)

```
qqnorm(ls,main = "Normal Q-Q Plot")
```



BMD Example(Cont.)

```
> t.test(l$)
```

two-sided test

One Sample t-test

data: l\$

t = -2.6003, df = 40, p-value = 0.01299 alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

-0.063720981 -0.007986336

sample estimates:

mean of x

-0.03585366

```
> t.test(l$,mu=0,alternative="less")
```

If we are interested in only alternatives on one sided test and know that $\mu < 0$, we may use one-side test

One Sample t-test

data: l\$

t = -2.6003, df = 40, p-value = 0.006494

alternative hypothesis: true mean is less than 0

95 percent confidence interval:

-Inf -0.01263611

sample estimates:

mean of x

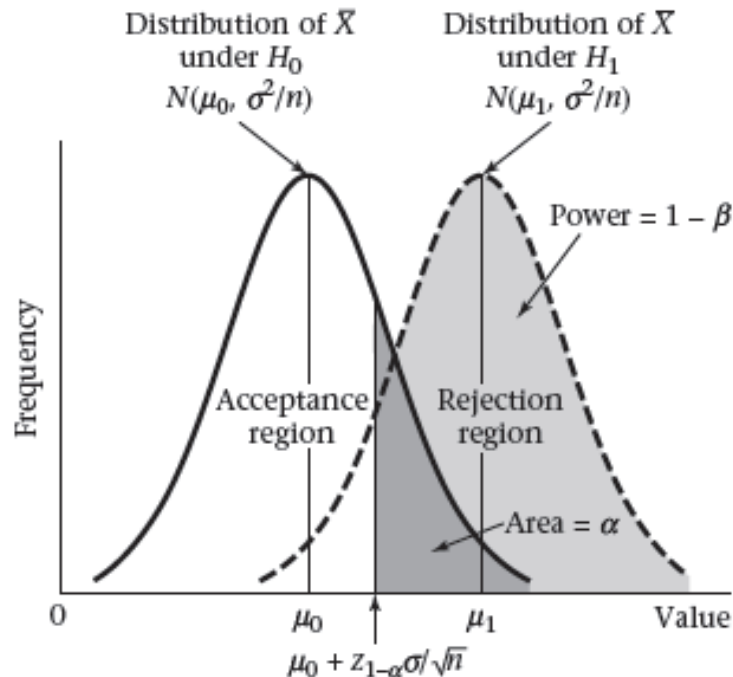
-0.03585366

more efficient

Power of a test

- Power is the probability of rejecting the null hypothesis when it is false. Note that $\text{power} = 1 - \beta$
- The calculation of power is used to plan a study, usually **before** any data have been obtained

Illustration of power for the one-sample test for the mean of a normal distribution with known variance ($\mu_1 > \mu_0$)



Notes

- Consider our previous example involving the familial aggregation of cardiovascular risk factors

$$H_0: \mu = 175 \quad \text{versus} \quad H_1: \mu > 175$$

- Then power is

$$P \left(\frac{\bar{X} - 175}{s/\sqrt{n}} > t_{1-\alpha, n-1} \middle| \mu = \mu_1 \right)$$

- Note that this is a function that depends on the specific value of μ_1 !
(But the nature of the test does not depend on the value chosen for μ_1 provided that μ_1 is more than 175 mg/dL)
- Notice as μ_1 approaches 175 the power approaches α .

Calculating power

- Assume that n is large and that we know σ

$$\begin{aligned} 1 - \beta &= P\left(\frac{\bar{X} - 175}{\sigma/\sqrt{n}} > z_{1-\alpha} \middle| \mu = \mu_1\right) \\ &= P\left(\frac{\bar{X} - \mu_1 + \mu_1 - 175}{\sigma/\sqrt{n}} > z_{1-\alpha} \middle| \mu = \mu_1\right) \\ &= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > z_{1-\alpha} - \frac{\mu_1 - 175}{\sigma/\sqrt{n}} \middle| \mu = \mu_1\right) \\ &= P\left(Z > z_{1-\alpha} - \frac{\mu_1 - 175}{\sigma/\sqrt{n}} \middle| \mu = \mu_1\right) \end{aligned}$$

Cont.

- Suppose that we wanted to detect an increase in the average cholesterol level of at least 175 mg/dL.
- Assume normality and that the sample in question will have a standard deviation of 10 ; Also assume that the significance level $\alpha = 0.05$.
- what would be the power if we took a sample size of 16?
- $Z_{1-\alpha} = 1.645$ and $\frac{\mu_1 - 175}{\sigma/\sqrt{n}} = \frac{5}{10/\sqrt{16}} = 2$
- $P(Z > 1.645 - 2) = P(Z > -0.355) = 64\%$

```
> library("pwr")  
> pwr.norm.test(d=0.5,n=16,alt="greater")
```

Mean power calculation for normal distribution with known variance

```
      d = 0.5  
      n = 16  
sig.level = 0.05  
  power = 0.63876  
alternative = greater
```

Power for the T test

- What about if we don't know σ ?
- Consider calculating power for a student's T test
- The power is

$$P\left(\frac{\bar{X} - 175}{S/\sqrt{n}} > t_{1-\alpha, n-1} \mid \mu = \mu_1\right)$$

and we have
$$\frac{\bar{X} - 175}{S/\sqrt{n}} = \frac{\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - 175}{\sigma/\sqrt{n}}}{S/\sigma} \sim t_{n-1, ncp = \frac{\mu_1 - 175}{\sigma/\sqrt{n}}}$$

Continued

- Let $\sigma = 10$ and $\mu_1 - \mu_2 = 5$

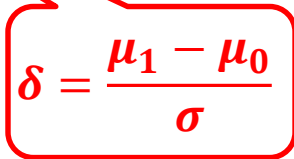
```
> power.t.test(n = 16, delta = 5/10,  
+             type = "one.sample", alt = "one.sided")
```

One-sample t test power calculation

```
          n = 16  
    delta = 0.5  
      sd = 1  
sig.level = 0.05  
  power = 0.6040329  
alternative = one.sided
```

Power analysis

- The calculation for $H_1: \mu < \mu_0$ is similar
- For $H_1: \mu \neq \mu_0$ calculate the one sided power using $\alpha/2$ (this is only approximately right, it excludes the probability of getting a large test statistic in the opposite direction of the truth)
- Power goes up as n gets larger
- Power goes up as the **effect size δ** increases, i.e., μ_1 gets further away from μ_0
- Power is better for one-sided tests.


$$\delta = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample-Size Determination

- For planning purposes, we frequently need some idea of an appropriate sample size for investigation **before** a study actually begins.
- Consider a study of the effect of a calcium-channel-blocking agent on heart rate for patients with unstable angina (Example 7.37).
- Suppose we want at least **80% power** for detecting a significant difference if the effect of the drug is to change mean heart rate by 5 beats per minute over 48 hours in either direction and $\sigma = 10$ beats per minute.
- How many patients should be enrolled in such a study?
- Typical values for the desired power are 80%, 90%, ..., and so forth.

- Test statistic $z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} = \sqrt{n} \frac{\bar{X} - \mu_1}{\sigma} + \sqrt{n} \frac{\mu_1 - \mu_0}{\sigma} \Rightarrow n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\delta^2}$
- For example: $\alpha = 0.05, \beta = 0.20, \delta = (180 - 175)/10 = 0.5$

$$\Rightarrow n = \frac{(Z_{0.975} + Z_{0.80})^2}{0.5^2} = \frac{(1.96 + 0.84)^2}{0.5^2} \approx 31.4$$
- Thus $n = 31.4$, or 32 people, would be needed.

```
> sample.size <- function(delta,alpha=0.05,beta){
+ n <- (qnorm(1-alpha/2)+qnorm(1-beta))^2/delta^2
+ list(sample.size=n)
+ }
> sample.size(delta=0.5,beta=0.2)
$sample.size
[1] 31.39552
```

```
> pwr.norm.test(d=0.5,power=0.8,alt="two.sided")
```

Mean power calculation for normal distribution with known variance

```
      d = 0.5
      n = 31.39544
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

```
> pwr.t.test(d=0.5,power=0.8,alt="two.sided",type="one.sample")
```

One-sample t test power calculation

```
      n = 33.36713
      d = 0.5
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

One-sided alternative

```
> pwr.norm.test(d=0.5,power=0.8,alt="greater")
```

Mean power calculation for normal distribution with known variance

```
      d = 0.5  
      n = 24.73022  
sig.level = 0.05  
  power = 0.8  
alternative = greater
```

```
> pwr.t.test(d=0.5,power=0.8,alt="greater",type="one.sample")
```

One-sample t test power calculation

```
      n = 26.13753  
      d = 0.5  
sig.level = 0.05  
  power = 0.8  
alternative = greater
```


Based on CI width

Find the minimum sample size needed to estimate cholesterol levels in the offspring if we require that the two-sided 95% CI for μ be no wider than 5 mg/dL and the sample standard deviation equals 10 mg/dL.

z test statistic (approximate)

$$n = 4z_{1-\alpha/2}^2 s^2 / L^2$$

$$n = \frac{4(z_{.975})^2 (10)^2}{5^2} = 61.5$$

Thus 62 patients need to be studied.

t test statistic

$$n = 4t_{\textcolor{red}{n-1}, 1-\alpha/2}^2 s^2 / L^2$$

There isn't explicit expression for n . Some algorithm will be used to find an appropriate n .
64 patients need to be studied

```
> ## calculation of the sample size based on CI width with z test statistic
```

```
> sam.size.CI <- function(sigma, width, alpha=0.05){
```

```
+ n <- 4*qnorm(1-alpha/2)^2*sigma^2/width^2
```

```
+ list(method = "CI width",sample.size=n)
```

```
+ }
```

```
> sam.size.CI(sigma=10,width=5)
```

```
$method [1]
```

```
"CI width"
```

```
$sample.size
```

```
[1] 61.46334
```

```
> ## calculation of the sample size based on CI width with t test statistic
```

```
> sam <- function(n){ n-4*qt(p=1-alpha/2,df=n-1)^2*sigma^2/width^2}
```

```
> samsize.CI.t <- function(sigma, width, alpha=0.05){
```

```
+ r <- uniroot(sam, interval=c(2,100))
```

```
+ list(method = "CI width with t test statistic",sample.size=r$root)
```

```
+ }
```

```
> samsize.CI.t(10,5)
```

```
$method
```

```
[1] "CI width with t test statistic"
```

```
$sample.size
```

```
[1] 63.8979
```

One-sample test for the Variance

- The χ^2 test for $H_0: \sigma^2 = \sigma_0^2$ versus
 - $H_1: \sigma^2 \neq \sigma_0^2$
- Test statistic
 - $X^2 = (n - 1)s^2 / \sigma^2$
- Reject the null hypothesis when
 - $X^2 < \chi_{n-1, \alpha/2}^2$ or $X^2 > \chi_{n-1, 1-\alpha/2}^2$
- Consider concerning the variability of blood-pressure measurements taken on an Arteriosonde machine. Assume $\sigma^2 = 35$ from the standard cuff and $s^2 = 8.178, n = 10$.
- Then the test statistic is given by $X^2 = 2.103$ and its p-value equals .021

Hypothesis tests for binomial proportions

Motivation example

- We were interested in the effect of having a family history of breast cancer on the incidence of breast cancer. Suppose that 400 of the 10,000 women ages 50–54 sampled whose mothers had breast cancer had breast cancer themselves at some time in their lives.
- Given large studies, assume the prevalence rate of breast cancer for U.S. women in this age group is about 2%.
- The question is: How compatible is the sample rate of 4% with a population rate of 2%?

Hypothesis tests for binomial proportions

- Consider testing $H_0: p = p_0$ for a binomial proportion
- The **score** test statistic

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

follows a normal distribution for large n

- This test performs better than the **Wald** test

$$\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

Example

- In our previous example, we want to test the hypothesis $H_0: p = .02$ vs. $H_1: p \neq .02$.

- $\hat{p} = 400/10,000 = .04$

- Test statistic (Score)

$$\frac{.04 - .02}{\sqrt{.02 * .98/10,000}} = 14.3$$

$$\text{p-value} = 2 * P(Z > 14.3) < .001, \text{ REJECT}$$

- Similarly, using the Wald test, $\text{p-value} = 2 * P(Z > 10.2) < .001, \text{ REJECT}$

Exact binomial tests

- Suppose that only 10 women were sampled and 4 of them had breast cancer at some time in their lives.
- Consider calculating an exact P-value
- What's the probability, under the null hypothesis, of getting evidence as extreme or more extreme than we obtained?

$$p = 2P(X \geq 4) = 2 \sum_{k=4}^{10} \binom{10}{k} \times .02^k \times .98^{10-k} \approx 0$$

```
> binom.test(4,10,.2,alt = "two.sided")
```


Hypothesis tests for the Poisson Distribution

Motivation example

- Many studies have looked at possible health hazards faced by rubber workers.
- In one such study, a group of 8418 white male workers ages 40–84 (either active or retired) on January 1, 1964, were followed for 10 years for various mortality outcomes .
- Their mortality rates were then compared with U.S. white male mortality rates in 1968.
- In one of the reported findings, 4 deaths due to Hodgkin's disease were observed compared with 3.3 deaths expected from U.S. mortality rates.
- Is this difference significant?

Hypothesis tests for Poisson proportions

- Consider testing $H_0: \mu = \mu_0$ for a Poisson distribution
- The large-sample test
 - Test statistic: $X^2 = \frac{(x - \mu_0)^2}{\mu_0}$
which follows a χ_1^2 distribution under H_0
- The exact test
 - `poisson.test(4, T=1, r=3.3)`