

Confidence interval for two-sample problem

Recall the model

$Y_{11}, \dots, Y_{1n_1} \stackrel{iid}{\sim} N(\mu_1, \sigma^2)$ independent of $Y_{21}, \dots, Y_{2n_2} \stackrel{iid}{\sim} N(\mu_2, \sigma^2)$.

Exactly, we have

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \sim N \left(\mu_1 - \mu_2, \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right] \right).$$

One could plug in the MLE $\hat{\sigma}^2$ for σ^2 above and get an approximate large sample CI.

The exact approach takes the variability in estimating σ^2 by $\hat{\sigma}^2$ into account, but rather uses the unbiased estimator

$$S_{pooled}^2 = \frac{\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2},$$

instead of the MLE $\hat{\sigma}^2$.

Then

$$T = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} - (\mu_1 - \mu_2)}{\text{se}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})} \sim t_{n_1+n_2-2},$$

where $\text{se}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})$ estimates $(\text{Var}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}))^{1/2}$ and is given by $S_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Quantiles $q_{0.025}$ and $q_{0.975}$ from a $t_{n_1+n_2-2}$

distribution are found such that for $T \sim t_{n_1+n_2-2}$,

$P(q_{0.025} \leq T \leq q_{0.975}) = 0.95$. Then a 95% CI for $\mu_1 - \mu_2$ is given by

$$(\bar{y}_{1\bullet} - \bar{y}_{2\bullet} + q_{0.025}\text{se}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}), \bar{y}_{1\bullet} - \bar{y}_{2\bullet} + q_{0.975}\text{se}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})).$$

We'll let R do the work for us.

```
> skull <- c(english,celt)
> group <- c(rep("English",length(english)),rep("Celt",length(celt)))
> group <- factor(group)
> t.test(skull~group,var.equal=T)
```

Two Sample t-test

```
data: skull by group t = -7.6952, df = 32, p-value = 9.003e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.91906 -11.58094
```

With 95% confidence we estimate that modern Englishmen have head breads between 11.6 and 19.9 *mm* larger than their Celtic ancestors, on average. Note that R gives a 95% CI for $\mu_2 - \mu_1$ not $\mu_1 - \mu_2$.

Hypothesis testing for non-normal data

Testing the exponential distribution: leukemia data

The normal-based procedures are powerful, but cannot be used on every data set.

For example, the lifetimes in weeks of $n = 33$ patients who died of acute myelogenous leukemia are 156, 65, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 5, 1, 1, 65, 56, 65, 17, 7, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43. On the next slide a histogram of the data along with the best fitting normal distribution (using $\hat{\mu} = \bar{y}_{33} = 40.9$ and $\hat{\sigma} = \sqrt{s_{33}^2} = 46.0$) shows a normal distribution provides very poor fit to these data.

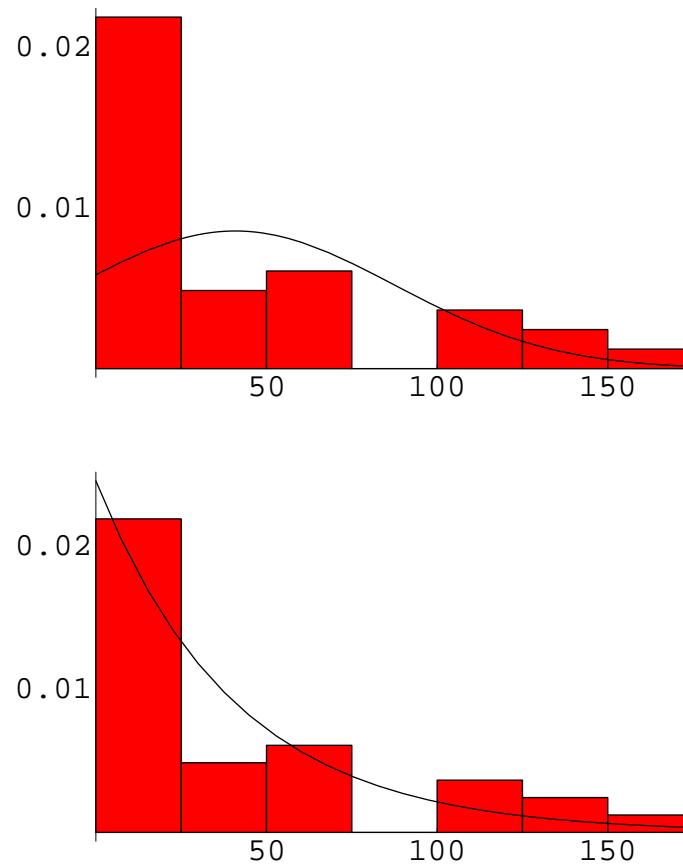


Figure 1: Histogram of leukemia survival times and MLE fits of normal and exponential models.

Instead, we will assume the data arises from an exponential distribution

$$Y_1, \dots, Y_{33} \stackrel{iid}{\sim} \exp(\theta).$$

Note then that each individual has a mean lifetime of $E(Y_i) = 1/\theta$. The MLE and the MOM of θ is $\hat{\theta} = 1/\bar{Y}_n$. Recall the large sample result $\hat{\theta} \stackrel{\bullet}{\sim} N(\theta, \theta^2/n)$ for exponential data. Here, $n = 33$ so $\hat{\theta} = 1/\bar{Y}_{33}$.

We estimate θ by $\hat{\theta} = 1/\bar{y}_{33} = 1/40.9 = 0.0245$. We don't know the approximate variance of $\hat{\theta}$, given by θ^2/n , but we can estimate it using the MLE as $\hat{\theta}^2/n = 0.0245^2/33 = 1.82 \times 10^{-5}$ and so before the lifetimes were actually observed we approximate the distribution of $\hat{\theta}$ based on $n = 33$ observations by

$$\hat{\theta} \stackrel{\bullet}{\sim} N(\theta, 1.82 \times 10^{-5}).$$

This is subtle. Before the data are collected, $\hat{\theta}$ is a *random function of the data*. After data are collected we actually see $\hat{\theta} = 0.0245$. We use this to estimate the variance of the *random estimator* $\hat{\theta}$ before the survival times are collected.

We wish to test

$$H_0 : E(Y_i) \geq 52 \text{ versus } H_1 : E(Y_i) < 52.$$

In terms of θ this is

$$H_0 : 1/\theta \geq 52 \text{ versus } H_1 : 1/\theta < 52,$$

or equivalently

$$H_0 : \theta \leq \frac{1}{52} = 0.0192 \text{ versus } H_1 : \theta > \frac{1}{52} = 0.0192.$$

We have cast the hypothesis of interest $H_0 : E(Y_i) \geq 52$ in terms of the population parameter θ .

We want to show $H_1 : \theta > 0.0192$; the null hypothesis is $H_0 : \theta \leq 0.0192$. The p -value is the probability of seeing a *random* $\hat{\theta}$ from an identical experiment that is larger than what we saw, $\hat{\theta} = 0.0245$ if $\theta = 0.0192$.

As before we compute the p -value as

$$\begin{aligned} p\text{-value} &= P(\hat{\theta} > 0.0245) \text{ given that } \theta = 0.0192 \\ &= P\left(\frac{\hat{\theta} - 0.0192}{\sqrt{0.0245^2/33}} > \frac{0.0245 - 0.0192}{\sqrt{0.0245^2/33}}\right) \\ &= P(Z > 1.229) = 0.11. \end{aligned}$$

This is a small p -value, but is it “small enough” to reject $H_0 : E(Y_i) \geq 52$? Not at the $\alpha = 0.05$ or even $\alpha = 0.1$ levels.

One sample proportion

Consider estimation and testing for a population proportion, e.g. the proportion π of French skiers that develop a cold.

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Bern}(\pi).$$

Recall that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. The parameter π is the unknown population proportion of individuals with the attribute of interest (e.g. those that develop a cold, or perhaps the proportion of Democrats in Washington county).

The MLE and MOM of π is $\hat{\pi} = \bar{Y}_n$, the number in the sample with the attribute of interest divided by the total sampled n .

since $E(Y_i) = \pi$ and $\text{Var}(Y_i) = \pi(1 - \pi)$, the Central Limit Theorem tells us

$$\hat{\pi} = \bar{Y}_n \overset{\bullet}{\sim} N \left(\pi, \frac{\pi(1 - \pi)}{n} \right).$$

We don't know the variance $\pi(1 - \pi)/n$ but we can estimate it by plugging in $\hat{\pi}$ for π yielding the *standard error* of $\hat{\pi} = \bar{Y}_n$:

$$\text{se}(\hat{\pi}) = \sqrt{\frac{\bar{y}_n(1 - \bar{y}_n)}{n}}.$$

A common test is $H_0 : \pi = \pi_0$ versus $H_0 : \pi \neq \pi_0$ where π_0 is a known test value. The p -value is computed as before, the probability of seeing a *sample proportion* \bar{Y}_n as far away or even further away from π_0 assuming $\pi = \pi_0$ is true.

$$\begin{aligned}
p\text{-value} &= P(|\bar{Y}_n - \pi_0| \geq |\bar{y}_n - \pi_0|) \text{ when } \pi = \pi_0 \\
&= P\left(\left|\frac{\bar{Y}_n - \pi_0}{\text{se}(\hat{\pi})}\right| \geq \left|\frac{\bar{y}_n - \pi_0}{\text{se}(\hat{\pi})}\right|\right) \text{ when } \pi = \pi_0 \\
&\approx P(|Z| \geq |z_0|),
\end{aligned}$$

where $z_0 = \frac{\bar{y}_n - \pi_0}{\text{se}(\hat{\pi})}$ is the test statistic computed from the observed sample and $Z \sim N(0, 1)$.

For the French skier example, the sample proportion that developed a cold was $\bar{y}_{139} = 17/139 = 0.1223$. This is our estimate of π , the proportional of all French skiers that develop colds after taking Vitamin C. Also, $\text{se}(\hat{\pi}) = \sqrt{0.1223(1 - 0.1223)/139} = 0.02779$.

To test $H_0 : \pi = 0.1$, form

$z_0 = (\bar{y}_{139} - \pi_0)/\text{se}(\hat{\pi}) = (0.1223 - 0.1)/0.02779 = 0.802$. The p -value for the test is thus $P(|Z| \geq 0.802) = 2\Phi(-0.802) = 0.422$. We cannot reject H_0 at any reasonable significance level.

To construct a 95% CI for π we start with

$P(-1.96 \leq Z \leq 1.96) = 0.95$ where $Z = (\bar{Y}_n - \pi)/\text{se}(\hat{\pi})$ and work backwards to

$$P(\bar{Y}_n - 1.96\text{se}(\hat{\pi}) \leq \pi \leq \bar{Y}_n + 1.96\text{se}(\hat{\pi})) \approx 0.95.$$

So a 95% CI for π is given by $(\bar{y}_n - 1.96\text{se}(\hat{\pi}), \bar{y}_n + 1.96\text{se}(\hat{\pi}))$. For the French skier data this is

$$(0.1223 - 1.96(0.02779), 0.1223 + 1.96(0.02779)) = (0.068, 0.177).$$

We are 95% ‘confident’ that the true proportion of French skiers that develop colds while taking this dose of ascorbic acid is between 6.8% and 17.7%.

The following built-in R code gets a slightly different p -value and CI:

```
> prop.test(17,139,0.1,correct=FALSE)
```

```
1-sample proportions test without continuity correction
```

```
data: 17 out of 139, null probability 0.1 X-squared = 0.7682, df
=1, p-value = 0.3808 alternative hypothesis: true p is not equal to
0.1 95 percent confidence interval:
```

```
0.07777872 0.18714057
```

```
sample estimates:
```

```
p
```

```
0.1223022
```

There is a package called `binom` that has various methods for constructing CI's for π . I installed and loaded the package into R.

```
> help(binom.confint)
```

```
> binom.confint(19,139)
```

yields the following 11 confidence intervals...

	method	x	n	mean	lower	upper
1	agresti-coull	17	139	0.1223022	0.07686825	0.1880510
2	asymptotic	17	139	0.1223022	0.06783556	0.1767688
3	bayes	17	139	0.1250000	0.07264917	0.1804032
4	cloglog	17	139	0.1223022	0.07447311	0.1826936
5	exact	17	139	0.1223022	0.07288577	0.1885839
6	logit	17	139	0.1223022	0.07740004	0.1879468
7	probit	17	139	0.1223022	0.07604074	0.1854223
8	profile	17	139	0.1223022	0.07488056	0.1835065
9	lrt	17	139	0.1223022	0.07487822	0.1835066
10	prop.test	17	139	0.1223022	0.07493561	0.1912843
11	wilson	17	139	0.1223022	0.07777872	0.1871406

We see our CI derived using large sample normal theory (i.e. the CLT) is given by “asymptotic” and the built-in R CI from `prop.test` is “Wilson.” The Wilson estimate is given by $\hat{\pi} \pm 1.96\text{se}(\hat{\pi})$ where, instead, “2” is added to each of the number of successes and the number of failures:

$$\hat{\pi} = \frac{\sum_{i=1}^n y_i + 2}{n + 4} \text{ and } \text{se}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n + 4}}.$$

Testing proportions in two samples

Similar to the two-sample normal model, we can compare two proportions within two subgroups of subjects. The full “French skier” data looked at a second group of 140 French skiers who took a placebo. The full data can be placed in a two-by-two contingency table:

	Vitamin C	placebo
cold	17	31
no cold	122	109
total	139	140

Now we have two groups:

$$Y_{11}, \dots, Y_{1n_1} \stackrel{iid}{\sim} \text{Bern}(\pi_1),$$

independent of

$$Y_{21}, \dots, Y_{2n_2} \stackrel{iid}{\sim} \text{Bern}(\pi_2).$$

We are interested in the difference in proportions $\pi_1 - \pi_2$. Properties of normal distributions and the CLT give us

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \stackrel{\bullet}{\sim} N \left(\pi_1 - \pi_2, \frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2} \right).$$

This can be used to construct a CI for $\pi_1 - \pi_2$. Similarly, a hypothesis test for $H_0 : \pi_1 - \pi_2 = 0$ can be constructed. We'll skip the details but they're straightforward.

Let's test for no difference between the probability of catching a cold over the two week study period between the Vitamin C and placebo groups.

```
> colds <- c(17,31)
> total <- c(139,140)
> prop.test(colds,total)
```

2-sample test for equality of proportions with continuity correction

```
data: colds out of total X-squared = 4.1407, df = 1, p-value =
0.04186 alternative hypothesis: two.sided 95 percent confidence
interval:
```

```
-0.194027721 -0.004225105
```

```
sample estimates:
```

```
prop 1    prop 2
0.1223022 0.2214286
```

Let π_1 be the probability of catching a cold while taking Vitamin C over the study period. Let π_2 be the probability of catching a cold taking placebo.

We reject $H_0 : \pi_1 = \pi_2$ at the 5% level in favor of $H_0 : \pi_1 \neq \pi_2$. There is a statistically significant difference in the probability of catching a cold when taking Vitamin C versus placebo. We would estimate that the probability of catching a cold is between 0.4% and 19.4% greater when taking placebo versus vitamin C with 95% confidence.

We might *a priori* be looking in a particular direction to start, i.e. $H_0 : \pi_1 \geq \pi_2$ versus $H_1 : \pi_1 < \pi_2$. Typing `prop.test(colds,total,alternative="less")` yields a *p*-value of 0.02093 and a 95% upper bound on $\pi_1 - \pi_2$ of -0.0183 .

Cervical dysplasia and HPV

This example is courtesy Prof. E. Bedrick at U.N.M. A case-control study was designed to examine risk factors for cervical dysplasia (Becker et al., 1994). The women in the study were patients at U.N.M. clinics aged 18 to 40. There were $n_1 = 175$ cases with cervical dysplasia and $n_2 = 308$ controls without. Each woman was tested for human papilloma virus (HPV):

	dysplasia	no dysplasia
HPV+	164	130
HPV-	11	178
total	175	308

Let π_d be the probability of HPV among the cases and π_n be the probability of dysplasia among the controls.

```
> positive <- c(164,130)
> total <- c(175,308)
> prop.test(positive,total)
```

2-sample test for equality of proportions with continuity correction

```
data: positive out of total X-squared = 122.1411, df = 1, p-value <
2.2e-16 alternative hypothesis: two.sided 95 percent confidence
interval:
 0.4447407 0.5853892
sample estimates:
  prop 1    prop 2
0.9371429 0.4220779
```

We reject $H_0 : \pi_d = \pi_n$ at any reasonable significance level, including $\alpha = 0.0001$. There is a strong statistical association between the presence of dysplasia and the presence of HPV.

In the above output, what does **X-squared = 122.1411** mean?

Recall that for large sample tests we form a test statistic Z_0 that has a standard normal distribution under the null hypothesis.

For any $Z_0 \sim N(0, 1)$, Z_0^2 has a chi-squared distribution with one *df*, written $Z_0^2 \sim \chi_1^2$. R denotes Z_0^2 as **X-squared**. The *p*-value is obtained as

$$p - \text{value} = P(Z_0^2 \geq z_0^2),$$

where z_0 is the observed test statistic.

The χ_{df}^2 distribution is special case of the gamma distribution.

For the leukemia data, in formulating and testing the hypothesis, we assumed exponential data. Do not confuse this assumption with the large sample result that uses normality. We assume the *data* are exponential and estimate the parameter θ . When n is large, $\hat{\theta}$ is approximately normal, not the data.

In the exponential and Bernoulli examples, normality, either exact or large sample, was used to solve the problem. In all cases the variance of the estimator was a *nuisance parameter* and estimated from the data in an intelligent way. This is a very common course of action in statistical hypothesis testing.

An alternative is to plug in the *hypothesized* value θ_0 when computing the variance of the sampling distribution of $\hat{\theta}$. This yields the *score test* and will be discussed later.