

Testing the mean in normal data, Case II: σ unknown

Recall that when n is large $\sigma \approx \hat{\sigma}$, the MLE value. Then the “ z -test” provides a very good approximate hypothesis test. However, when n is small to moderate in size (i.e. say $n < 20$ or so) this approximation can be improved on.

The approximation works by taking into account the variability in estimating σ^2 by $\hat{\sigma}^2$. Let $T_0 = (\bar{Y}_n - \mu_0) / \sqrt{S_n^2 / (n - 1)}$ where $S_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is the MLE of the variance σ^2 .

This is the same as the previous test statistic with σ known except we’re using S^2 instead of σ^2 and dividing by $n - 1$ instead of n . This test statistic has a t distribution with $n - 1$ degrees of freedom, written

$$T_0 = \frac{\bar{Y} - \mu_0}{\sqrt{S_n^2 / (n - 1)}} \sim t_{n-1}.$$

The t_{df} distribution has fatter tails than the $N(0, 1)$ reflecting the added variability in estimating σ^2 by S_n^2 . However, when $df > 30$ the t_{df} density is essentially a $N(0, 1)$ density and there's no practical difference in using a t_{df} versus a $N(0, 1)$ – the p -values are almost exactly the same. In fact, $t_n \xrightarrow{P} N(0, 1)$. See Figure 6.1 on p. 194.

The two tests (where the variability of estimating σ by $\hat{\sigma}$ is or isn't accounted for) are called z -tests and t -tests. In R, the t -test is implemented in `t.test()`. Type `help(t.test)` to see a list of arguments. We get a larger p value compared to the z -test ($p = 0.0002$):

```
> t.test(cal, alternative="greater", mu=100)
      One Sample t-test
data:  cal t = 3.4308, df = 16, p-value = 0.001715 alternative
hypothesis: true mean is greater than 100

95 percent confidence interval:
 109.2156      Inf
sample estimates: mean of x
 118.7647
```

Significance tests

Recall our model

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

If σ is known, and $H_0 : \mu = \mu_0$ is true then

$$Z_0 = (\bar{Y}_n - \mu_0) / \sqrt{\sigma^2/n} \sim N(0, 1).$$

If σ is unknown and $H_0 : \mu = \mu_0$ is true then

$$T_0 = (\bar{Y}_n - \mu_0) / \sqrt{S_n^2/(n-1)} \sim t_{n-1}.$$

Either way, if H_0 is true then the test statistic Z_0 or T_0 is one observation from a known, unimodal distribution that is symmetric about zero.

We consider σ unknown and consider the T_0 statistic only. The three possible alternatives are:

1. If the alternative is $H_1 : \mu > \mu_0$, or equivalently $H_1 : \mu - \mu_0 > 0$, then the numerator of T_0 , $\bar{Y}_n - \mu_0$, estimates what is on the left of this inequality. Thus $T_0 > 0$ provides evidence that the null H_0 is false and the alternative H_1 is true.

We reject H_0 if $\bar{Y}_n - \mu_0$ is much larger than zero relative to its standard deviation, which is $\sqrt{\text{Var}(\bar{Y}_n - \mu_0)} = \sqrt{\sigma^2/n}$. This is estimated by the *standard error* of \bar{Y}_n , $\text{se}(\bar{Y}_n) = \sqrt{S_n^2/(n-1)}$. (In fact, $E(S_n^2/(n-1)) = \sigma^2/n$).

We *normalize* our estimate of $\mu - \mu_0$, given by $\bar{Y}_n - \mu_0$, by an estimate of how variable it is, $\sqrt{S_n^2/(n-1)}$, giving us the test statistic

$$T_0 = \frac{\bar{Y}_n - \mu_0}{\sqrt{S_n^2/(n-1)}}.$$

When $T_0 > 0$ we have evidence in favor of $H_1 : \mu > \mu_0$, so when T_0 is large and positive we reject $H_0 : \mu \leq \mu_0$. We need to quantify how big is “big.”

We will formally reject H_0 if $T_0 > t_{crit}$, a “critical value” that defines a *decision rule* of whether we reject or not. We reject H_0 in favor of $H_1 : \mu > \mu_0$ if $T_0 > t_{crit}$ and we accept H_0 if $T_0 < t_{crit}$.

The question “how big does t_{crit} have to be to disprove H_0 ?” is answered by fixing the probability of wrongly rejecting the null (committing a Type I error) at some small α , typically $\alpha = 0.05$, i.e.

$$P(\text{reject } H_0 | H_0 \text{ is true}) = \alpha.$$

This translates into a statement that we can solve for t_{crit} :

$$P(T_0 > t_{crit} | \mu = \mu_0) = \alpha.$$

A picture helps.

The critical value t_{crit} is thus the value such that

$P(T_0 > t_{crit}) = \alpha$ where $T_0 \sim t_{n-1}$, and can be obtained from statistical packages or from the table in the back of your book.

Note that t_{crit} is different for different sample sizes n and different error rates α .

Comments:

The test statistic T_0 provides evidence against the null hypothesis $H_0 : \mu \leq \mu_0$ when it's positive. The larger it is, the more unlikely H_0 is.

We construct a formal test by rejecting H_0 when $T_0 > t_{crit}$ where t_{crit} is positive.

Even when H_0 is true we can still see large, positive T_0 . We let t_{crit} be defined by fixing the probability of seeing a T_0 larger than t_{crit} to be α when H_0 is really true.

α is called the “significance level” of the hypothesis test.

2. If the alternative is $H_1 : \mu < \mu_0$, or equivalently $H_1 : \mu - \mu_0 < 0$, then $T_0 < 0$ provides evidence that the null is false and the alternative is true. How small T_0 has to be is given by fixing the maximum Type I error to be α ,

$$P(T_0 < t_{crit} | \mu = \mu_0) = \alpha.$$

The critical value, t_{crit} , is such that $P(T_0 < t_{crit}) = \alpha$ where $T_0 \sim t_{n-1}$.

3. If the alternative is $H_1 : \mu \neq \mu_0$, or equivalently $H_1 : \mu - \mu_0 \neq 0$, then T_0 away from zero in *either* direction provides evidence that the null is false and the alternative is true. How large $|T_0|$ has to be is given by fixing the Type I error to be α ,

$$P(|T_0| > |t_{crit}| | \mu = \mu_0) = \alpha.$$

where $T_0 \sim t_{n-1}$.

Let's consider the p -value for the alternative $H_1 : \mu \neq \mu_0$. Before data are collected, $T_0 \sim t_{n-1}$. After data are collected we see $T_0 = t_0$, the test statistic. Seeing t_0 away from zero gives evidence that $H_0 : \mu = \mu_0$ is not true and p -value = $P(|T_0| > |t_0|)$ is the probability of seeing a T_0 as far away or even further away from zero as t_0 .

Now a significance test at level α picks the t_{crit} such that $P(|T_0| > t_{crit}) = \alpha$. If we see $|t_0| > t_{crit}$ then we reject H_0 as false while controlling the Type I error rate at α . These two paragraphs imply the following:

$$p\text{-value} \leq \alpha \quad \text{if and only if} \quad \text{Reject } H_0 \text{ at level } \alpha.$$

This statement is true regardless of the direction of the alternative (one-sided or two-sided). Thus one only need report the p -value and a significance test at any level α can be immediately performed.

Recall for the calorie content of poultry hot dogs,

```
> t.test(cal,alternative="greater",mu=100)
      One Sample t-test
data:  cal t = 3.4308, df = 16, p-value = 0.001715 alternative
hypothesis: true mean is greater than 100
```

The p -value is 0.001715. We reject at the $\alpha = 0.05$ and $\alpha = 0.01$ significance levels, but not $\alpha = 0.001$.

Let's say in a paper we're writing that we state "...we reject the null hypothesis $\mu \leq 100$ at the 1% level." We are stating that we reject H_0 at $\alpha = 0.01$, i.e. that t_0 was far enough away from $\mu_0 = 100$ that by *rejecting* $\mu_0 \leq 100$ we only have a 1 in 100 chance of committing a type I error.

Hypothesis testing for the two-sample problem

The model is written

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

There are three unknown population parameters in the model

$\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$. Recall that the MLE's of μ_1 and μ_2 are the respective *sample* means $\bar{y}_{1\bullet} = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j}$ and $\bar{y}_{2\bullet} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j}$.

We are often interesting in testing $H_0 : \mu_1 - \mu_2 = 0$, or equivalently

$H_0 : \mu_1 = \mu_2$ versus one of the alternatives $H_1 : \mu_1 - \mu_2 \neq 0$

($\mu_1 \neq \mu_2$); $H_1 : \mu_1 - \mu_2 > 0$ ($\mu_1 > \mu_2$); or $H_1 : \mu_1 - \mu_2 < 0$ ($\mu_1 < \mu_2$).

Note that $\bar{Y}_{1\bullet} \sim N(\mu_1, \sigma^2/n_1)$ independent of $\bar{Y}_{2\bullet} \sim N(\mu_2, \sigma^2/n_2)$, so using properties of the normal distribution

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

One could “plug in” the MLE for σ^2 above,

$\hat{\sigma}^2 = [\sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \mu_2)^2] / (n_1 + n_2)$ and do an approximate test based on the $N(0, 1)$ distribution, i.e. if

$H_0 : \mu_1 - \mu_2 = 0$ is true then from the previous line

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\left[\frac{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2}{n_1 + n_2}\right] \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \stackrel{\bullet}{\sim} N(0, 1).$$

An 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 $\hat{\sigma}^2$.

As before in the one-sample case, we form a test statistic which has a simple distribution when $H_0 : \mu_1 = \mu_2$ is true

$$T_0 = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\text{se}(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet})} \sim t_{n_1+n_2-2},$$

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 $S_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. p -values are computed as usual.

For the head breadth data we obtain a p -value of about 0.00000001. If indeed $H_0 : \mu_1 = \mu_2$ is true, then the probability of seeing the two sample means as far apart as we saw or even further is about one in one hundred million. Yes, it is possible, but highly unlikely. In fact, if we reject H_0 based on this p -value there is a one in one hundred million chance that we made a mistake, i.e. committed a type I error by rejecting H_0 when it's in fact true.

As far as significance testing goes, we reject at $\alpha = 0.05$, $\alpha = 0.01$, or even $\alpha = 0.001$.

```
> 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
```

Confidence intervals

We may be interested in providing a plausible range in which μ lies in the one-sample problem, or in which $\mu_1 - \mu_2$ lies in the two sample problem.

Confidence intervals – one sample problem

Let

$$Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

Then we know for the true unknown μ ,

$$\bar{Y}_n \sim N(\mu, \sigma^2/n),$$

and

$$T = \frac{\bar{Y}_n - \mu}{\sqrt{S_n^2/(n-1)}} = \frac{\bar{Y}_n - \mu}{\text{se}(\bar{Y}_n)} \sim t_{n-1}.$$

Let's say that we know $n = 20$. A t_{19} density looks like:

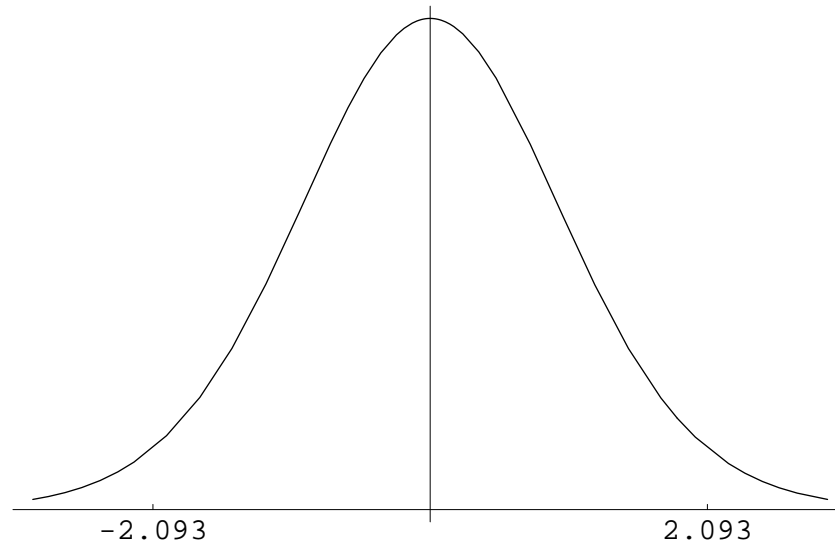


Figure 1: Density for $T \sim t_{19}$ random variable.

Let $T \sim t_{19}$. We can compute that $P(-2.093 \leq T \leq 2.093) = 0.95$.

These are the 0.025 and 0.975 quantiles of T .

A 95% probability interval for μ is derived:

$$\begin{aligned} 0.95 &= P(-2.093 \leq T \leq 2.093) \\ &= P\left(-2.093 \leq \frac{\bar{Y}_n - \mu}{\sqrt{S_n^2/(n-1)}} \leq 2.093\right) \\ &= P\left(-2.093\sqrt{S_n^2/(n-1)} \leq \bar{Y}_n - \mu \leq 2.093\sqrt{S_n^2/(n-1)}\right) \\ &= P\left(\bar{Y}_n - 2.093\sqrt{S_n^2/(n-1)} \leq \mu \leq \bar{Y}_n + 2.093\sqrt{S_n^2/(n-1)}\right). \end{aligned}$$

This is true before we collect data and actually see $\bar{Y}_n = \bar{y}_n$ and $S_n^2 = s_n^2$. After we see these two statistics we can still form the interval, but it is no longer random. Since it is not random we talk about “confidence.”

This is for $n = 20$. For other sample sizes n , we will have different quantiles than -2.093 and 2.093 .

For the poultry hot dog data, we see MLE's $\bar{y}_{17} = 118.77$ and $s_{17}^2 = 478.65$. Let $T \sim t_{16}$; then $P(-2.120 \leq T \leq 2.120) = 0.95$ (there's a different t_{df} density for every df).

Then

$$(\bar{y}_n - 2.120\sqrt{s_n^2/(n-1)}, \bar{y}_n + 2.120\sqrt{s_n^2/(n-1)})$$

becomes

$$(118.77 - 2.120\sqrt{478.65/16}, 118.77 + 2.120\sqrt{478.65/16})$$

equalling (107.17, 130.36).

We are 95% “confident” that μ is between 107.2 and 130.4 calories.

Important point: there *is no randomness* here. Either $(107.17, 130.36)$ covers the unknown μ or it doesn't; we don't know. However, before we collected the data Y_1, \dots, Y_{17} there was a 95% chance that the resulting interval would cover μ .

We can think of performing an experiment where we collect 10000 independent samples of size $n = 17$. We expect roughly 9500 of the resulting confidence intervals to cover μ and 500 to miss the mark.

The R function `t.test()` gives confidence intervals automatically. The confidence interval is the same regardless of what you specify for μ_0 . However, the CI *does* change depending on what you specify for the alternative. To get a 95% confidence interval (rather than a lower or upper bound), perform a two-sided test, the default in R.

```
> t.test(cal, mu=100)
```

```
One Sample t-test
```

```
data: cal t = 3.4308, df = 16, p-value = 0.00343 alternative
hypothesis: true mean is not equal to 100 95 percent confidence
interval:
 107.1698 130.3596
sample estimates: mean of x
 118.7647
```

Comments:

- Often introductory statistics textbooks have students compute several CI's by hand using sample statistics and quantiles from t_{n-1} distributions. As long as you understand the basic idea outlined above, we'll let R do the work for us.
- Remember that μ is fixed and unknown. The resulting CI either covers μ or it doesn't. Before we collect data Y_1, \dots, Y_n there is a 95% chance it *will* cover μ and a 5% chance it will not.
- We developed 95% confidence intervals, but you can also derive, e.g., 90% CI's or 99% CI's if you want. In R you'd specify `t.test(data, conf.level=0.99)`.

- You might not need an interval, but rather an upper or lower bound for μ . We won't discuss this, but it's very easy to get these in the same manner as for a two-sided interval. A lower bound is obtained `t.test(data, alternative="greater")` and an upper bound is obtained `t.test(data, alternative="less")`. These are also how you specify alternatives for hypothesis testing (but remember to include `mu=10` or whatever μ_0 is).
- Recall that under general circumstances, the elements $\hat{\theta}_1, \dots, \hat{\theta}_p$ of the MLE vector $\hat{\theta}$ are approximately normal with $\hat{\theta}_j \overset{\bullet}{\sim} N(\theta_j, \text{Var}(\hat{\theta}_j))$. The $\text{Var}(\hat{\theta}_j)$ is unknown but often estimated by the data as $\text{se}(\hat{\theta}_j)^2$. For $Z \sim N(0, 1)$, $P(-1.96 \leq Z \leq 1.96) = 0.95$ so an approximate 95% CI for θ is given by

$$(\hat{\theta} - 1.96\text{se}(\hat{\theta}_j), \hat{\theta} + 1.96\text{se}(\hat{\theta}_j)).$$

This is used by R, SAS, SPSS, etc. for a number of models.

Hypothesis testing and CI's in simple linear regression

The model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Usually, one is mainly interested in the slope β_1 .

Recall that exactly, $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$, where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

If we know σ^2 , then we know $\text{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ and we can pluck out the variance of $\hat{\beta}_1$ as the lower right entry.

You will show that the lower right entry of $(\mathbf{X}'\mathbf{X})^{-1}$ is $\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

However, we never know σ^2 . Instead we replace σ^2 by an unbiased estimate, $\hat{\sigma}_n^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2$.

So we define the *standard error*, an estimate of the standard deviation, of $\hat{\beta}_1$ as

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_n^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

If β_1 is the *true* value, then

$$T_0 = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2},$$

exactly. This forms the basis of a hypothesis test $H_0 : \beta_1 = 0$ and obtaining a CI for β_1 .

```

> score <- c(37.01,26.51,36.51,40.7,37.1,33.9,41.8,33.4,41.01,37.2,23.3,
+ 35.2,34.9,33.1,22.7,39.7,31.8,31.7,43.1,41.01)
> ses <- c(7.2,-11.71,12.32,14.28,6.31,6.16,12.7,-0.17,9.85,-0.05,
+ -12.86,0.92,4.77,-0.96,-16.04,10.62,2.66,-10.99,15.03,12.77)
> model.fit <- lm(score~ses)
> summary(model.fit)

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.32280	0.52800	63.11	< 2e-16 ***
ses	0.56033	0.05337	10.50	4.2e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- $se(\hat{\beta}_1) = 0.05337$.
- If $H_0 : \beta_1 = 0$ is true, then $t_0 = \frac{0.56033-0}{0.05337} = 10.5$ is a draw from a t_{18} distribution.
- The p -value for testing $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is $P(|t_{18}| > 10.5) = 0.0000000042$.
- R tells us this is significant at the $\alpha = 0.001$ level by placing “***” next to the p -value.

```
> confint(model.fit)
                2.5 %    97.5 %
(Intercept) 32.2135138 34.4320820
ses          0.4482012 0.6724497
```

These are 95% CI's. The CI for β_1 solves

$$0.95 = P(-2.101 \leq t_{18} \leq 2.101) = P\left(-2.101 \leq \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \leq 2.101\right)$$

to get

$$0.95 = P\left(\hat{\beta}_1 - 2.101\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 2.101\text{se}(\hat{\beta}_1)\right),$$

and plugs in the estimates $\hat{\beta}_1 = 0.56033$ and $\text{se}(\hat{\beta}_1) = 0.05337$.

If $T \sim t_{18}$ then $P(T \leq -2.101) = 0.025$ and $P(T \leq 2.101) = 0.975$; i.e. these are the 2.5% and 97.5% percentiles of a t_{18} distribution.