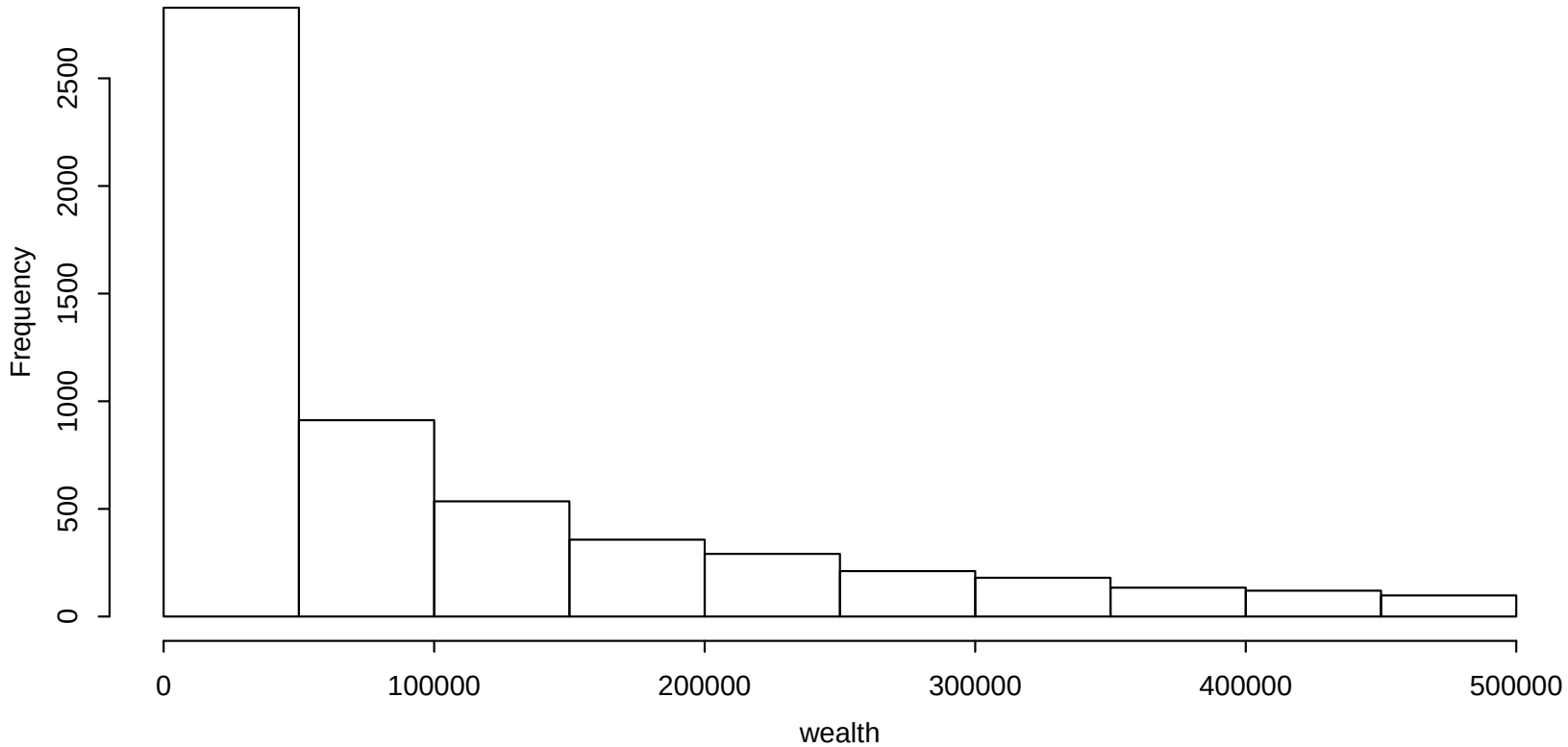
Sometimes T-dist doesn't fit I

Histogram of population wealth between 0 and 500000:

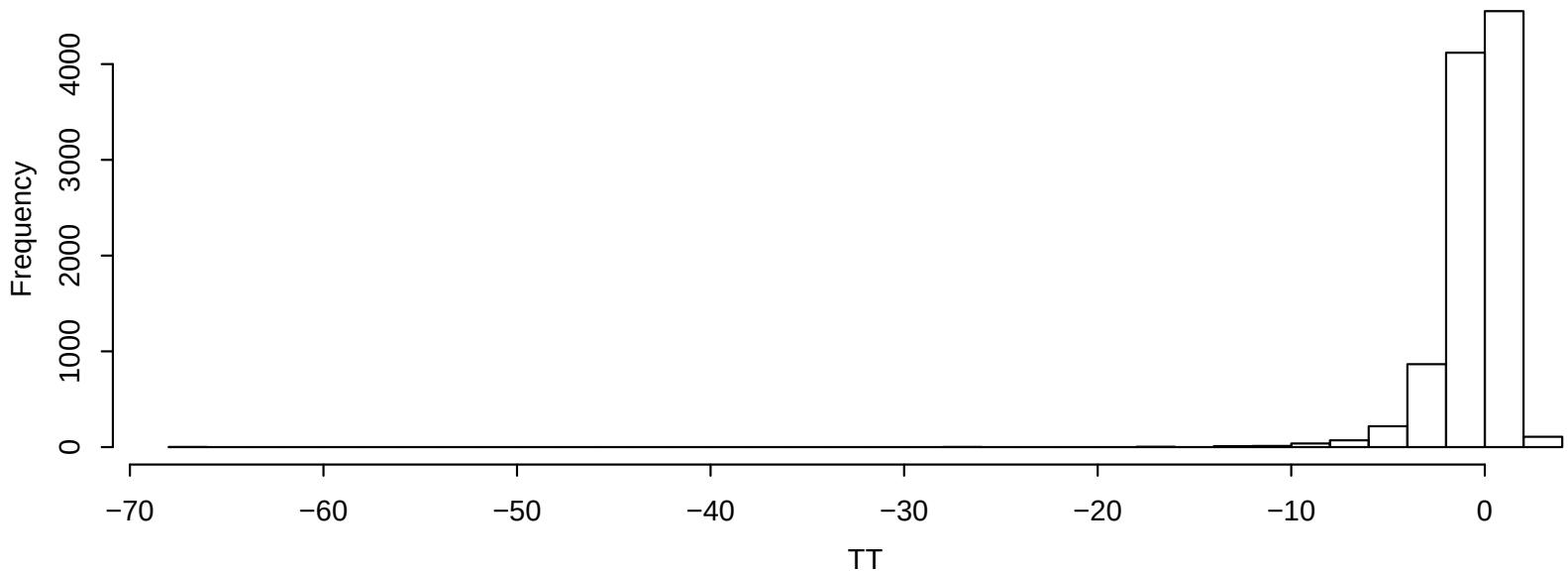


Mean is about 100000

Sometimes T-dist doesn't fit II

Histogram of T test-statistic for samples of size 10.

```
sampsize=10
wealths=replicate(10000,sample(wealth,sampsize))
ave=apply(wealths,2,mean);sds=apply(wealths,2,sd);
TT=(ave-MEAN)/(sds/sqrt(sampsize))
hist(TT,breaks=40,main="")
```



Histogram is very skewed.

Sometimes T-dist doesn't fit III

Imagine our sample was:

```
Y

 [1] 189500  55600  14800  22000   1000  48600  16600  56000
 [9]  19000 102105

c(mean(Y),sd(Y))

[1] 52520 56452

t=(mean(Y)-100000)/(sd(Y)/sqrt(10))
PY=2*(1-pt(abs(t),9))
c(t,PY)

[1] -2.65966  0.02606
```

Sometimes T-dist doesn't fit IV

We want to test the hypothesis that mean wealth is 100000 vs. the alternative that it is not.

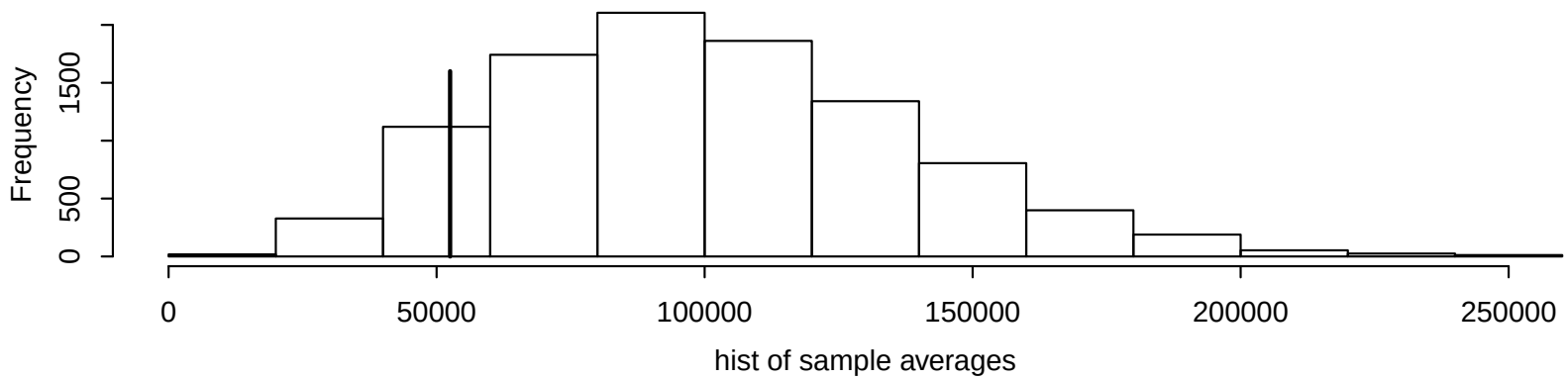
According to the T-test the p-value is 0.0261 - quite small.

Sometimes T-dist doesn't fit V

Since in this case we know the full population we can check what proportion of the 10000 samples of size 10 from the population have a mean that is further than 52520.5 from 100000 (above or below).

```
del=100000-my
mean(ave<100000-del | ave > 100000+del)
```

```
[1] 0.2046
```



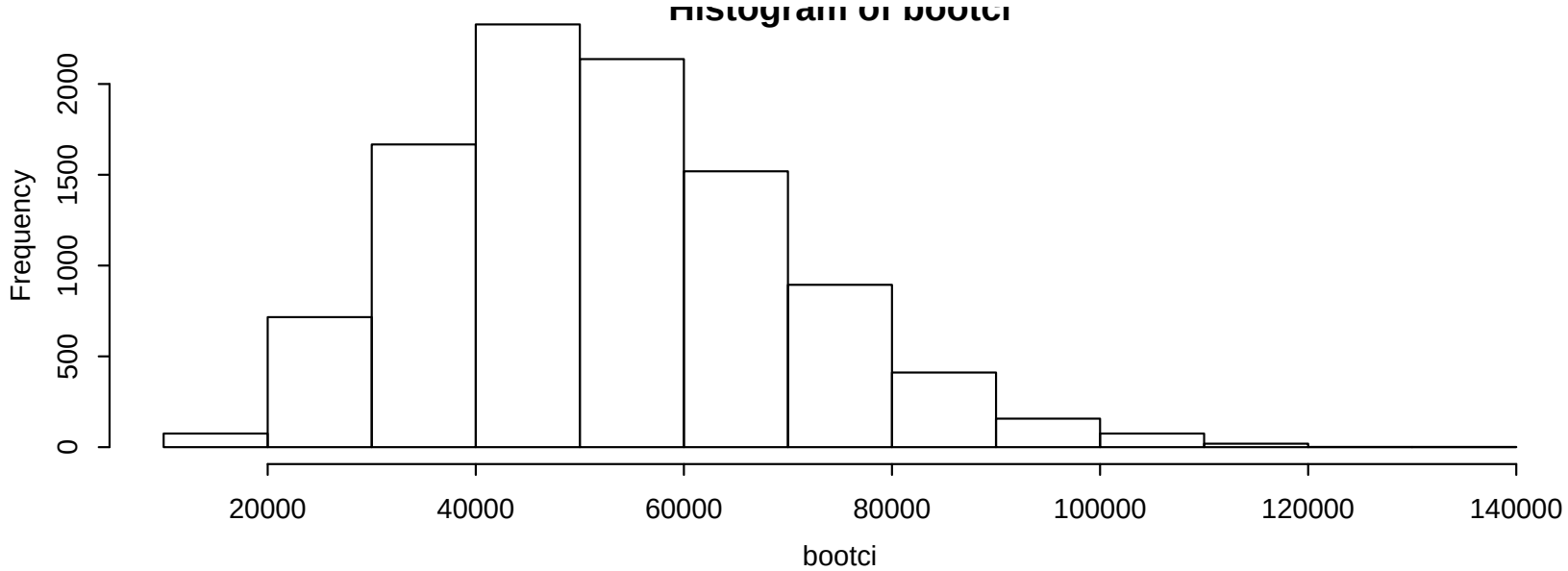
So the mean of 52520.5 is **not** so unlikely at all under the null.

Bootstrap I

In real life we only have **one** sample to work with, we can't sample multiple times from the population.
So we pretend Y describes the population. We can draw 10000 times from the sample Y and look at the distribution of means estimated from each of these samples, and find the $\alpha/2$ and $1-\alpha/2$ quantiles. If the histogram is skewed (as it is below) its an indication that the T-test may not be valid.

Bootstrap II

```
bootci=replicate(10000,mean(sample(Y,10,replace=TRUE)))
hist(bootci)
```




```
quantile(bootci,c(.025,.975))
```

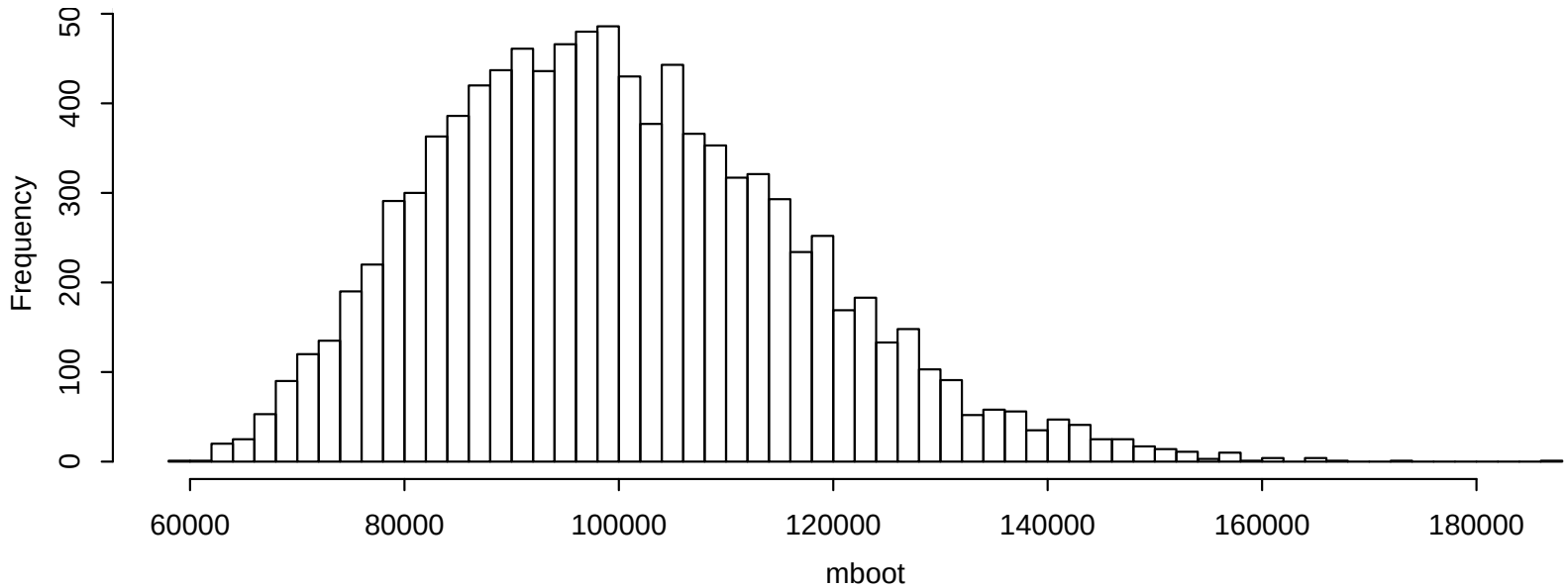
```
 2.5% 97.5%
23970 90191
```

Bootstrap III

For hypothesis testing shift sample to center it around the **hypothesized mean**, so at least the center of the samples will be the hypothesized mean. Then draw 10000 samples with replacement, compute the mean and look at the histogram.

```
my=mean(Y)
YH=Y-my+100000
boot=replicate(10000,sample(YH,10,replace=TRUE))
mboot=apply(boot,2,mean)
hist(mboot,main="",breaks=50)
lines(c(my,my),c(0,1000),lwd=2)
```

Bootstrap IV



We can use the multiple computed means to test the original statistic directly:

```
del=100000-my
mean(mboot<100000-del | mboot>100000+del)
```

```
[1] 0.0074
```

Bootstrap V

So according to the bootstrap sample the observed mean is very unlikely under the null...

The main conclusion should be that because the histogram of means is skewed **something is wrong** and there is no easy fix.

Inference for proportion I

Suppose we want to estimate the proportion p of some characteristic of a population, and we undertake the following procedure:

1. Draw a SRS of size n.
2. Record the number X of "successes" (those individuals having the characteristic).
3. Estimate the unknown true population proportion p with the sample proportion of successes $\hat{p} = \frac{X}{n}$

p is the mean of the population, but in this case it determines the $SD = \sqrt{p(1-p)}$. In the general case σ is not determined by μ. What is the sampling distribution of $\hat{p}$?

Inference for proportion II

If n is sufficiently large – i.e. if

$$np \geq 10 \text{ and } n(1-p) \geq 10,$$

then

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Thus, an approximate $(1-\alpha)$ CI for the population proportion p is given by

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is chosen so that $P(Z > z^*) = \alpha/2$ for $Z \sim N(0,1)$.

Inference for proportion III

What if we want to test whether $p = p_0$ for some fixed value p_0?
The null hypothesis is $p = p_0$, and under this hypothesis,

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$$

Notice that we are using a different value for the SD of $\hat{p}$ than was used for the CI. Since H_0 specifies a true value for p, the SD of $\hat{p}$ under H_0 is given by $\sqrt{\frac{p_0(1-p_0)}{n}}$
The p-values for this test are:

- $H_a : p > p_0 \qquad P(Z \geq z)$
- $H_a : p < p_0 \qquad P(Z \leq z)$
- $H_a : p \neq p_0 \qquad 2P(Z \geq |z|)$

for $Z \sim N(0,1)$.
Some care needs to be taken when p is very close to 0 or 1.

Inference for proportion IV

A random sample of 2700 California lawyers revealed only 1107 who felt that the ethical standards of most lawyers are high (*AP, Nov. 12, 1994*).

1. Does this provide strong evidence for concluding that fewer than 50% of all California lawyers feel this way?
2. What is a 90% confidence interval for the true proportion of California lawyers who feel that ethical standards are high?

Inference for proportion V

```
hp=1107/2700
Z=(hp-.5)/sqrt(.25/2700)
Z^2

[1] 87.48

prop.test(1107,2700,conf.level = .90,correct=FALSE)


1-sample proportions test without continuity
correction

data:  1107 out of 2700
X-squared = 87, df = 1, p-value <2e-16
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.3945 0.4257
sample estimates:
   p
0.41
```

Matched pairs t-test I

In a matched pairs study, there are 2 measurements taken on the same subject (or on 2 similar subjects). For example,

- 2 rats from the same litter
- before and after observations on the same subject
- adjacent plots on a field

To conduct statistical inference on such a sample, we analyze the *difference* using the one-sample procedures described above.

Matched pairs t-test II

Example: Diet and Weight

```
load(file="w.Rdata")
glimpse(weight)

Rows: 20
Columns: 4
$ Subject    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1...
$ weighta    <int> 187, 175, 158, 160, 130, 170, 165, 1...
$ weightb    <int> 160, 153, 150, 148, 127, 160, 150, 1...
$ difference <dbl> 27, 22, 8, 12, 3, 10, 15, -1, 10, 6,...

t=(mean(weight$difference)-0)/(sd(weight$difference)/sqrt(20))
p=1-pt(t,19)
c(t,p)

[1] 4.8842514 0.0000515
```

Matched pairs t-test III

To ascertain whether the diet reduces weight, we test

$$H_0 : \mu = 0 \qquad H_a : \mu > 0$$

where μ is the mean weight difference.

T-statistic: $t = \frac{9.35-0}{8.56/\sqrt{20}} = 4.88$

p-value: $p = P(t_{19} \geq 4.88) = 5.2 \times 10^{-5}$

Matched pairs t-test IV

Example: Income 2000 vs 2016

```
load(file="WealthIncome.Rdata")
Inc=na.omit(subset(Income,select=c("INCOME2000","INCOME2016")))
t.test(Inc$INCOME2000,Inc$INCOME2016,paired=TRUE)
```

```
Paired t-test

data:  Inc$INCOME2000 and Inc$INCOME2016
t = -15, df = 3900, p-value <2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25748 -19623
sample estimates:
mean of the differences
                 -22686
```

But... year 2000 dollar is not same is year 2016. So let's adjust:

```
cla=prod((1+COLA$X2[20:37]/100))
cla
```

```
[1] 1.469
```

Matched pairs t-test V

```
t.test(Inc$INCOME2000*cla,y=Inc$INCOME2016,paired=TRUE)
```

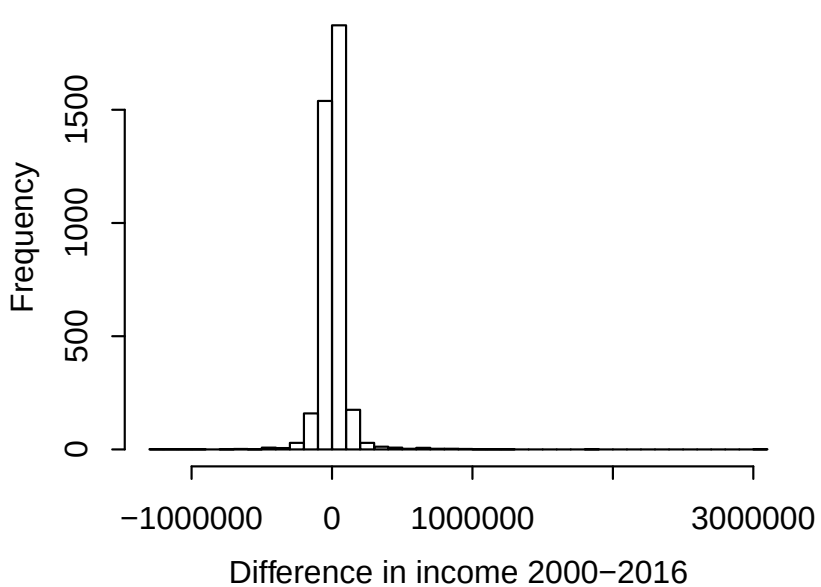
```
Paired t-test

data:  Inc$INCOME2000 * cla and Inc$INCOME2016
t = 3.8, df = 3900, p-value = 0.0001
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3489 10828
sample estimates:
mean of the differences
                   7159
```

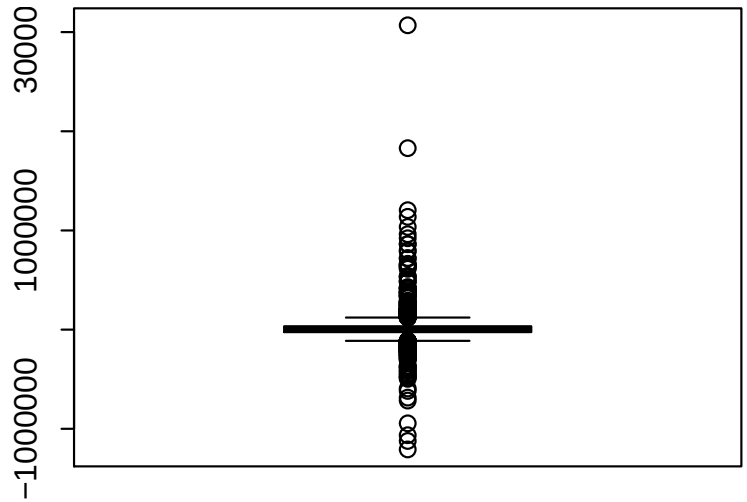
So actually, incomes have dropped on average during the past two decades.

Matched pairs t-test VI

But, we know that income and wealth data are highly skewed, so is this a valid test. Let's check the histogram of the **difference**

```
par(mfrow=c(1,2))
Incdiff=Inc$INCOME2000*cla-Inc$INCOME2016
hist(Incdiff,breaks=50,main="",xlab="Difference in income 2000-2016")
boxplot(Incdiff)
```

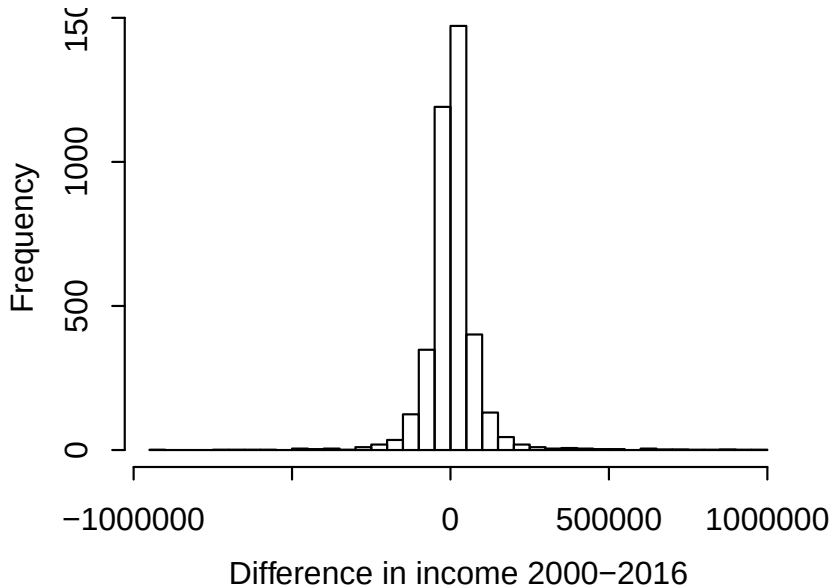


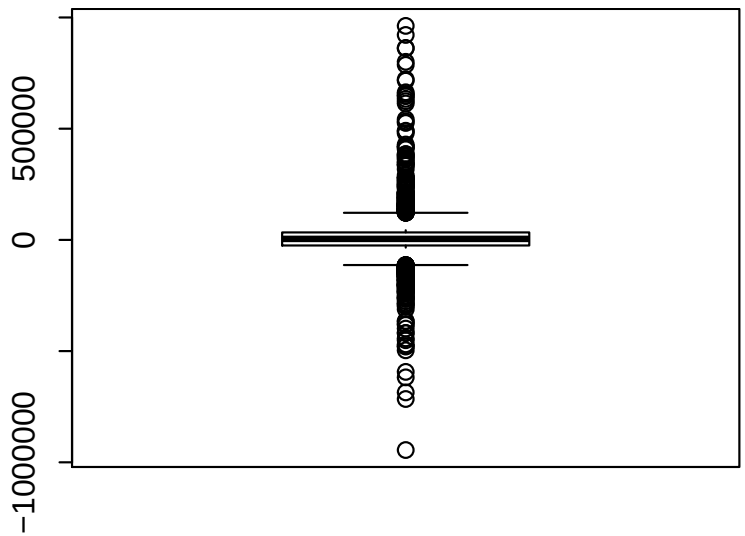


Looks nice an symmetric, but let's remove a couple of extreme outliers before testing:

Matched pairs t-test VII

```
par(mfrow=c(1,2))
Incdiffs=subset(Incdiff,Incdiff>-1000000 & Incdiff<1000000)
hist(Incdiffs,breaks=50,main="",xlab="Difference in income 2000-2016")
boxplot(Incdiffs)
```





Matched pairs t-test VIII

And now let's test again:

```
t.test(Incdiffs)
```

```
One Sample t-test

data:  Incdiffs
t = 4, df = 3900, p-value = 0.00006
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3034 8785
sample estimates:
mean of x
     5909
```

So... indeed real incomes have declined on average between 2000 and 2016.

Two sample problems I

- The goal of two-sample inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group (in addition to being independent of each other).

For example, Suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1)$ population and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2)$ population.
The first sample might be heights of male students and the second heights of female students.

We might test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$.

Two sample problems II

How is this different from the *matched pairs design*?

1. There is no matching of the units in two samples.
2. The two samples may be of different size.

Two sample problems III

Comparing Two Means when σ's are Known

Suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1)$ population (with sample mean $\bar{X}_1$) and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2)$ population (with sample mean $\bar{X}_2$). Suppose σ_1 and σ_2 are known.

The **two-sample z-statistic** is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Why the denominator? Since the two samples are independent, their averages are independent so:

$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Two sample problems IV

- A $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  where $z^* : P(Z > z^*) = \alpha/2$.

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0$$

  The p-value is calculated as before

Two sample problems V

Comparing Two Means with σ's Unknown

We define
$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1,i} - \bar{X}_1)^2, S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{2,i} - \bar{X}_2)^2.$

The **Two-sample t-statistic** is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \;\dot{\sim}\; t_\nu$$

The T statistic only has an **approximate** t_ν distribution with
$\nu = \frac{(w_1+w_2)^2}{w_1^2/(n_1-1)+w_2^2/(n_2-1)}$, $w_1 = S_1^2/n_1$, $w_2 = S_2^2/n_2$.

Two sample problems VI

- A $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) \pm t^* \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \text{ where } t^* : P(T_\nu > t^*) = \alpha/2.$$

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} t_\nu \text{ under } H_0$$

  The p-value is calculated as before.

Setting $\nu = \min(n_1 - 1, n_2 - 1)$ is simpler and yields a more conservative approximate procedure. That is, the CIs are longer than the true CI and p-values are larger than the true p-values.

Two sample problems VII

Pooled two-sample t procedures

In the previous procedure, we assumed that $\sigma_1 \neq \sigma_2$. What if we have reason to believe $\sigma_1 = \sigma_2 = \sigma$ (even though we don't know either value)?

We can gain information (i.e. power) by *pooling* the two samples together for estimating the variance:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

If the two populations are normal this is the exact distribution of T.

Two sample problems VIII

- A $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm t^* S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t^* : P(T_{n_1+n_2-2} > t^*) = \alpha/2$.

- To test the hypothesis $H_0 : \mu_1 = \mu_2$, we use

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \text{ under } H_0$$

The p-value is calculated as before.

Two sample problems IX

Weight gains (in kg) of babies from birth to age one year are measured. All babies weighed approximately the same at birth.

Group A	5	7	8	9	6	7	10	8	6
Group B	9	10	8	6	8	7	9		

Assume that the samples are randomly selected from independent normal populations. Is there any difference between the true means of the two groups?

i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known

ii) Assume σ_1 and σ_2 are unknown and unequal.

iii) Assume σ_1 and σ_2 are unknown but equal

State the hypothesis:

$$H_0 : \mu_1 = \mu_2 \qquad H_a : \mu_1 \neq \mu_2$$

Two sample problems X

Stats:

$$\bar{X}_1 = 7.33 \quad \bar{X}_2 = 8.14$$
$$S_1 = 1.58 \quad S_2 = 1.35$$
$$n_1 = 9 \qquad n_2 = 7$$

Two sample problems XI

i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known. Then, the two-sample z statistic is

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_1\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
$$= \frac{7.33 - 8.14}{1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.07$$

The two-sided p-value is

$$2P(Z \geq |z|) = 2P(Z \geq 1.07) = 0.28$$

where $Z \sim N(0, 1)$.
So there is no difference between the true population mean of these two group at the significance level 0.1.

Two sample problems XII

A 90% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X}_1 - \bar{X}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= (7.33 - 8.14) \pm 1.645 \times 1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}$$

$$= (-2.05\ ,\ 0.43)$$

As expected, the 90% confidence interval covers 0. Thus, we have 90% confidence that there is no difference between the true population means.

Two sample problems XIII

ii) Assume σ_1 and σ_2 are unknown and unequal. Then, the two-sample t statistic is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{7.33 - 8.14}{\sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}}} = -1.10$$

The two-sided p-value is

$$2P(T \geq |z|) = 2P(T \geq 1.10) = 0.31$$

where $T \sim t_6$.

Two sample problems XIV

A 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$
\begin{aligned}
(\bar{X}_1 - \bar{X}_2) \pm t^* \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} & \\
&= (7.33 - 8.14) \pm 1.94 \times \sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}} \\
&= (-2.23\ ,\ 0.61)
\end{aligned}
$$

where $P(|T| < t^*) = 0.90$. That is, $P(T > t^*) = 0.05$ or $t^* = t_{\nu,.05}$.

Two sample problems XV

iii) Assume σ_1 and σ_2 are unknown but equal.
The pooled two-sample estimator of σ is

$$\begin{aligned} S_p &= \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \\ &= \sqrt{\frac{(9-1)\times 1.58^2+(7-1)\times 1.35^2}{9+7-2}} \\ &= 1.54 \end{aligned}$$

Thus, the pooled two-sample t statistic is

$$t = \frac{\bar{X}_1-\bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{7.33-8.14}{1.54\sqrt{\frac{1}{9}+\frac{1}{7}}} = -1.04$$

Two sample problems XVI

The two-sided p-value is given by

$$2P(T \geq |t|) = 2P(T \geq 1.04) = 0.32 \qquad \text{where } T \sim t_{14}.$$

A 90% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\bar{X}_1 - \bar{X}_2) \pm t^* S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= (7.33 - 8.14) \pm 1.77 \times 1.54 \times \sqrt{\frac{1}{9} + \frac{1}{7}} \\ &= (-2.18\ ,\ 0.56)\end{aligned}$$

Where $P(|T| < t^*) = 0.90$. That is, $P(T > t^*) = 0.05$.