BRN Research Methods Workshop 2

2. Probability and Statistical Theory

Inhwan Ko (Univ. of Washington, Seattle)

July 22th, 2021

Contents

1. A Fundamental Problem to Causal Identification- Review
2. Probability Theory
   - Law of Large Numbers
   - Central Limit Theorem
3. Statistical Theory
   - Hypothesis Testing
4. Introduction to Linear Regression

1. A Fundamental Problem to Causal Identification-Review

$$\bar{Y}^t_{i\in T} - \bar{Y}^c_{i\in C} = [ATE] + [\text{baseline difference}] + (1-w)[\text{heterogenous treatment}]$$

1) Baseline difference occurs because the potential outcome without treatment can be different between the treatment and control groups. For instance, smokers may already have higher level of stress than non-smokers even if they did smoke from the first place. This is also called *pre-treatment bias*.

2) Heterogenous treatment can be produced because the treatment can interact with unobservable factors that are different between the treatment and control groups. For instance, if smokers have already higher level of stress than non-smokers, smoking makes stress level even higher and therefore further worsen their health status than it would worsen the health status of the non-smokers should they smoke. This is also called *post-treatment bias*.

1. A Fundamental Problem to Causal Identification-Review

Observation studies often use several control variables to reduce the bias of their SATE of interest. However, this is not a simple task.

Let z be a stress level which we will use as a control variable. If:

1) z both affects the treatment assignment and the observed outcome, then controlling for it mitigates *pre-treatment bias*. In this case, z is called a *confounder*.

2) z affects the observed outcome but also is affected by the treatment assignment, then controlling for it exacerbates *post-treatment bias.* In this case, z is called a *collider*.

Conclusion: Controlling for covariates is not always good!

1. A Fundamental Problem to Causal Identification-Review

- Random sampling matters less for estimating ATE
- Random treatment assignment is crucial; no correlation between treatment assignment and potential outcome is a sufficient condition for SATE to be a consistent estimator of ATE
- Controlling for covariates helps mitigate pre-treatment bias, but may worsen post-treatment bias
- Must have a good reason when controlling for covariates

1.1. SATE to ATE: Probabilitical and statistical perspective

Last week, what we've covered is an "econometric" perspective on a fundamental problem to causal identification. There's also a "statistical" perspective which we will cover today.

Key question: How to infer ATE (average treatment effect) from SATE (sample average treatment effect)?

Econometrician: Make treatment assignment uncorrelated with potential outcome (i.e., random treatment assignment).

Statistician: Make sample as large as possible.

cf) Josh Angrist: What's the Difference between Econometrics and Statistics?

1.1. SATE to ATE: Probabilitical and statistical perspective (cont.)

Econometricians are interested in "causal inference"- causal relationship between X and Y- using the logic of *counterfactual.*

Statisticans are interested in "statistical inference"- association between X and Y at a population level through sample- using the logic of *Neo-humean Regularity.*

This is why p-value should not be read as "explanatory power" or "causal effect"- they only tell us how confidence we are with this "statistical inference," not "causal inference."

2. Probability theory

In econometric perspective, the quantity of interest (QoI) was the treatment effect- which was calculated by the potential outcome given treatment minus the potential outcome given no treatment.

$$\hat{\bar{\delta}} = \bar{Y}^t_{i \in T} - \bar{Y}^c_{i \in C}$$

In probabilistic / statistical perspective, the quantity of interest (QoI) is the same, but instead calculated by the conditional mean of Y given X.

$$\hat{\bar{\delta}} = E(Y|X)$$

This can be generalized into cases where X is not only dichotomous but also continuous variable.

2.1. Can SATE be ATE? Revisited

Since we are using sample not population data, "can SATE be ATE?" type of question also holds in the statistics setting.

In econometric perspective, two sufficient conditions were $\bar{Y}^t_{i\in T} = \bar{Y}^t_{i\in C}$ *and* $\bar{Y}^c_{i\in C} = \bar{Y}^c_{i\in T}$.

In probabilistic (or frequentist) perspective, we need two theorems: (1) Law of Large Number and (2) Central Limit Theorem.

2.2. Law of Large Numbers

Let p be a probability of an event A. When an event A occurs r times out of n times of independent trials, for any real number $\epsilon > 0$:

$$lim_{n \to \infty} P(|\frac{r}{n} - p| > \epsilon) = 0$$

For instance, even though you (mathmatically) know that the probability of having 2 heads out of 3 coins is:

$$_3C_2(\frac{1}{2})^2(\frac{1}{2})^1 = \frac{3}{8}$$

But there is a chance that you will *always* have two heads when you throw coins only 10 times.

When you throw it, let say 100,000,000 times, the probability will converge into $\frac{3}{8}$.

2.3. Central Limit Theorem

Let's say your sample size is n. Or, you took n observations $(X_1, X_2, ..., X_n)$ from the population and calculated its mean $(\bar{X})$.

You do this k times for the same sample size n. Therefore, you'll have: $\bar{X}_1, \bar{X}_2, ..., \bar{X}_k$.

The probability distribution of $\bar{X}_k$ converges into *normal distribution* as $n \to \infty$.

There is a lot versions of its proof online, so search for it if you are more interested in!

This time I will execute the theorem with multiple simulations.

2.3. Central Limit Theorm (cont.)

Let's generate population data with 10,000 observations, which follow binomial distribution with a true probability of 0.3 with the size of 10. For instance, assume there are 10,000 batters in the world with the batting average of 0.3, and you observed how many hits they produce for each 10 times of hitting opportunity. Assign this to an object called "pop".

```
library(MASS)
set.seed(2021)

pop <- rbinom(10000, 10, 0.3)
pop[1:10]
```

```
##  [1] 3 4 4 2 3 4 3 2 4 6
```

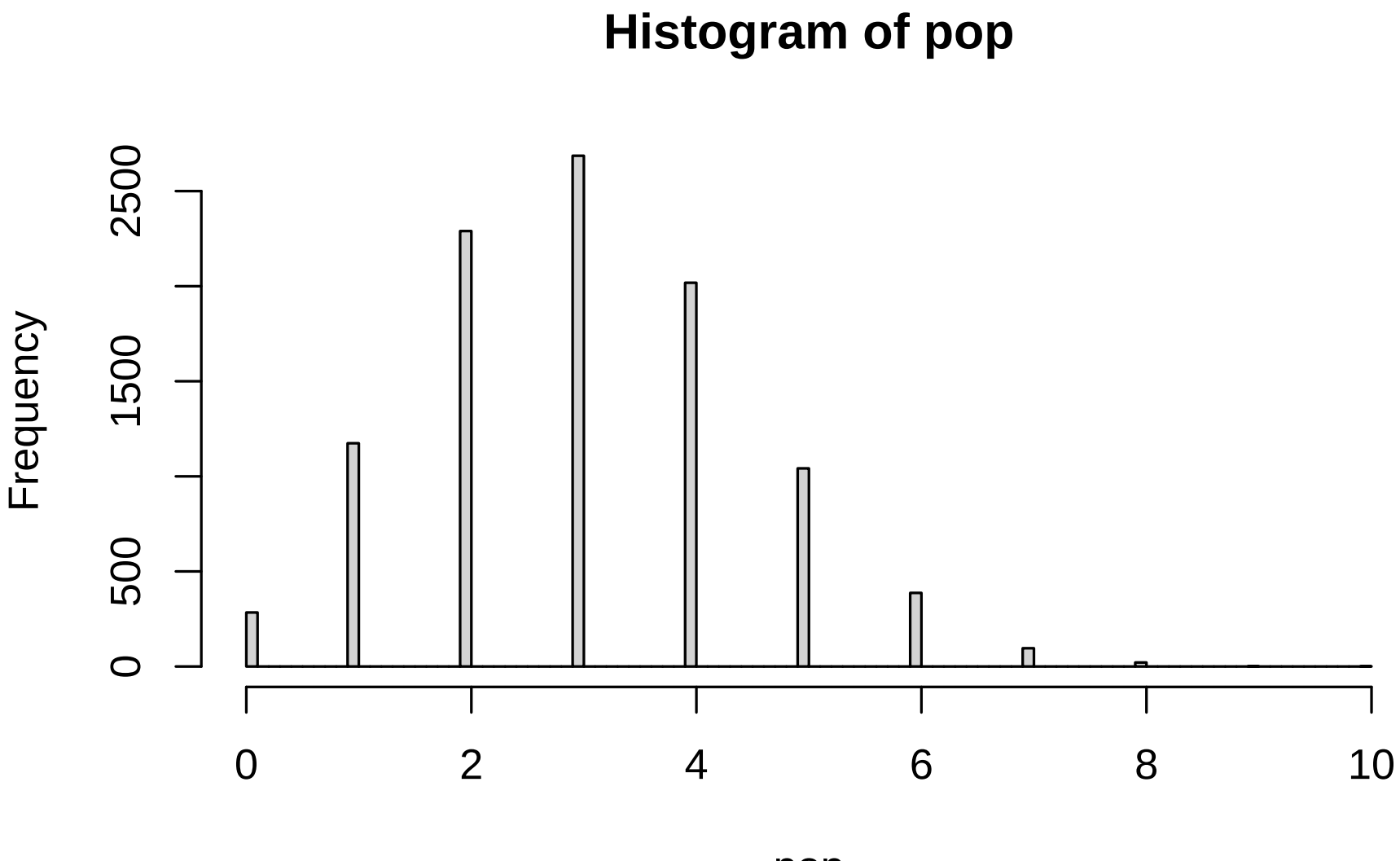
Let's take n observations from the population k times and make this a new function.

```
samplemean <- NULL

clt <- function(n, k){
  for (i in 1:k) {
  samplemean[i] <- mean(sample(pop, n, replace=T))
  }
  return(samplemean)
}
```

2.3. Central Limit Theorm (cont.)

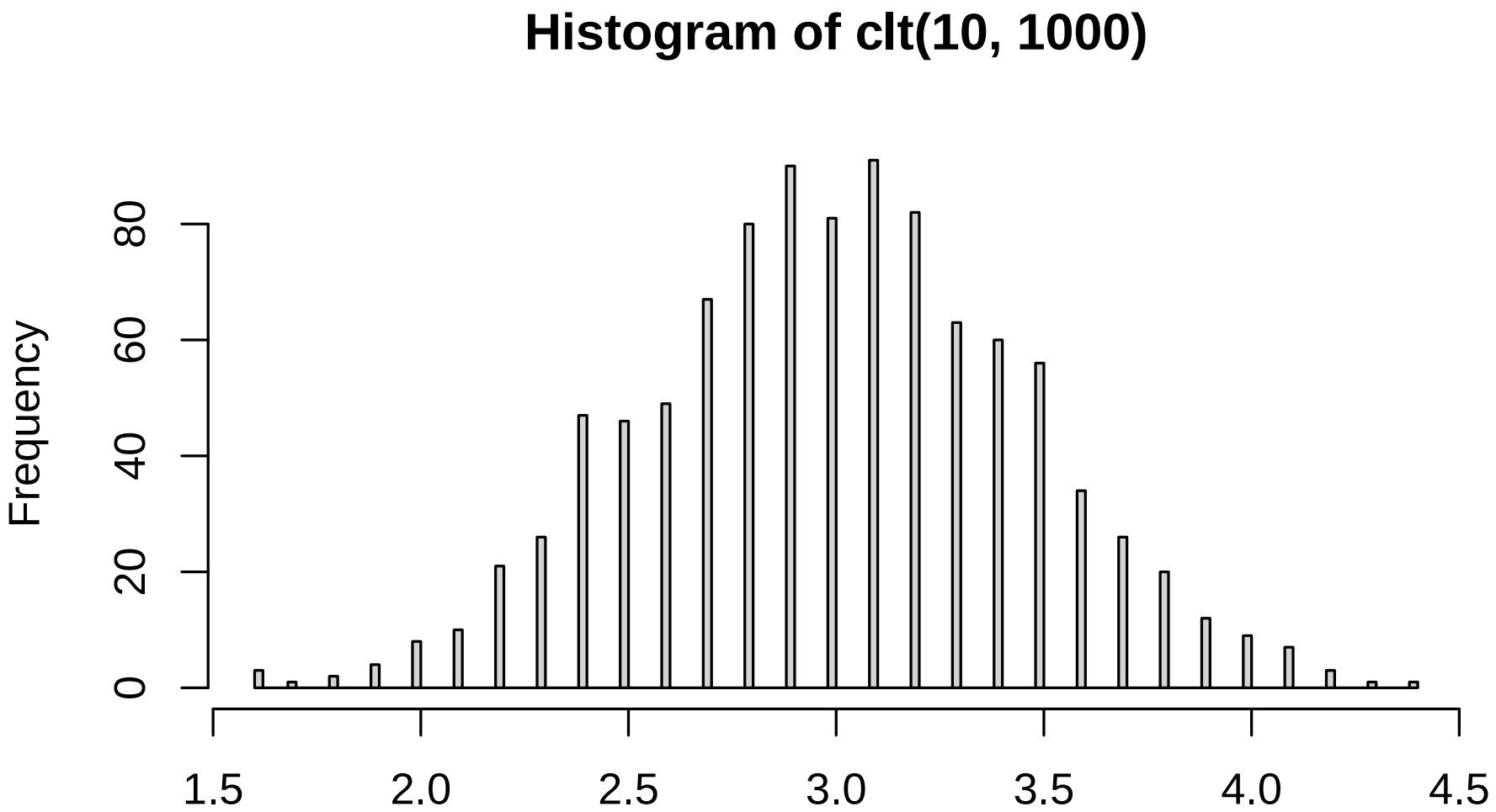
```
hist(pop, breaks=100)
```



2.3. Central Limit Theorem (cont.)

10 observations: does it look like a normal distribution?

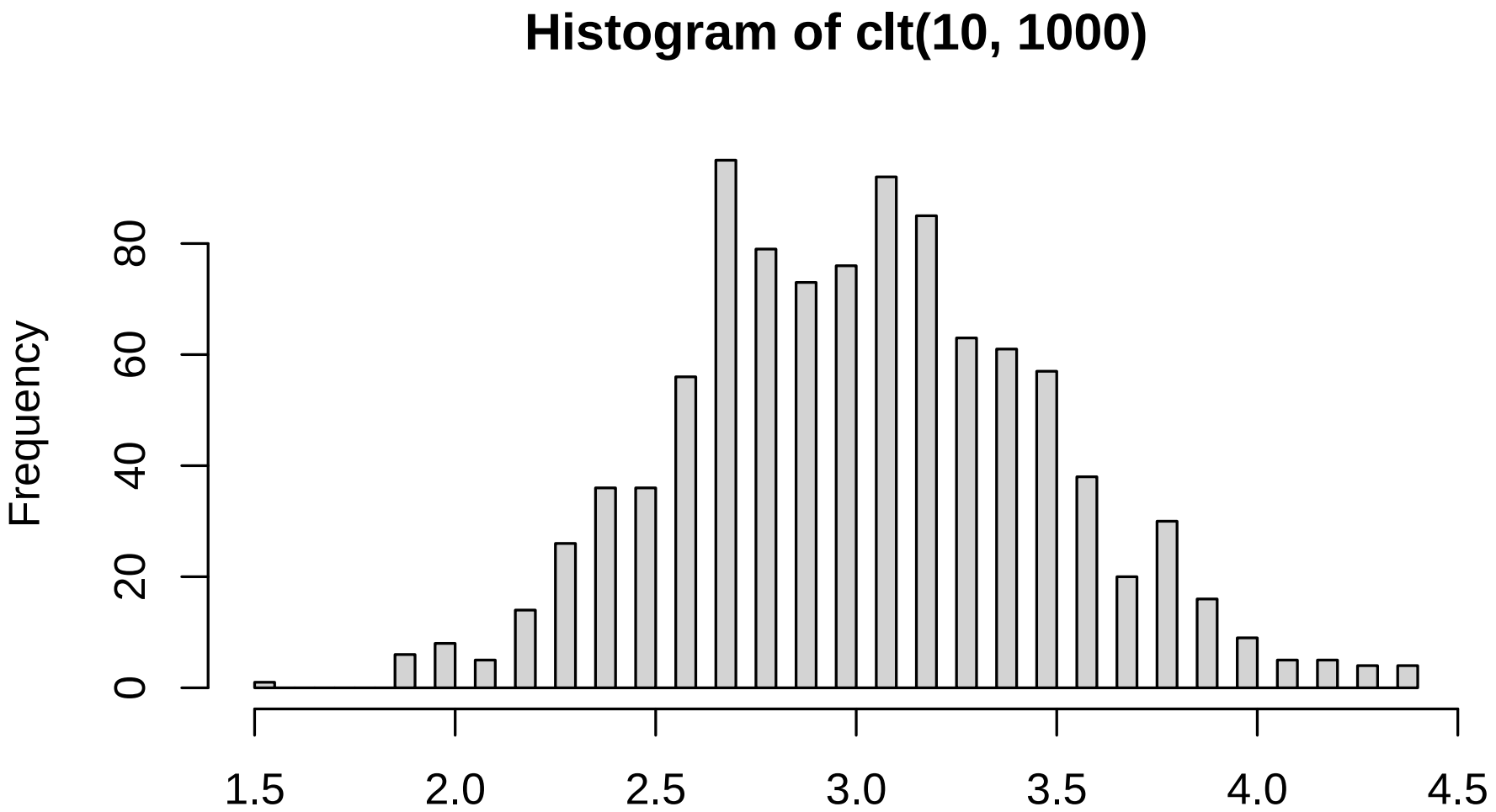
```
hist(clt(10,1000), breaks=100)
```



2.3. Central Limit Theorem (cont.)

20 observations: what about now?

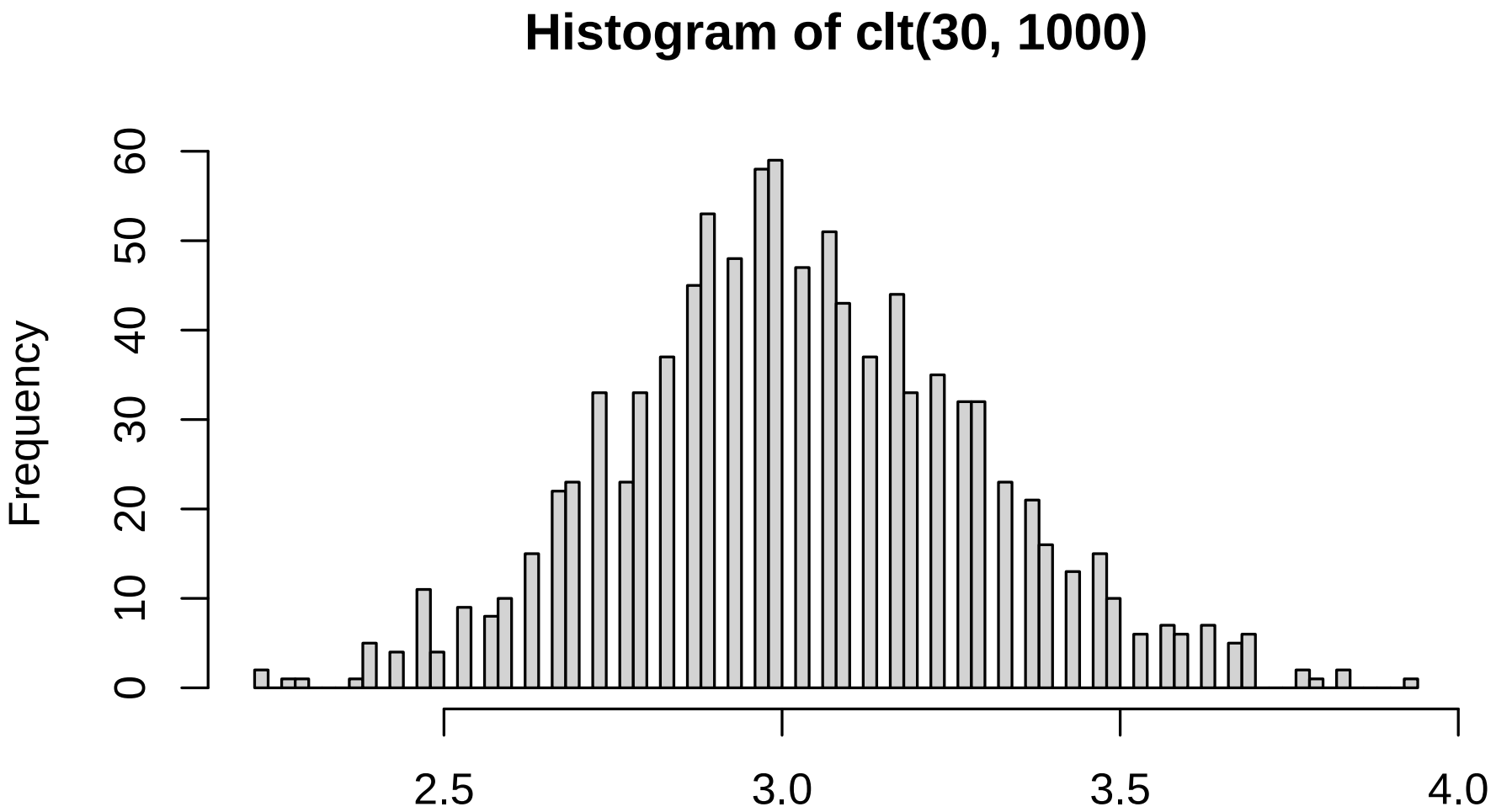
```
hist(clt(10,1000), breaks=100)
```



2.3. Central Limit Theorem (cont.)

30 observations?

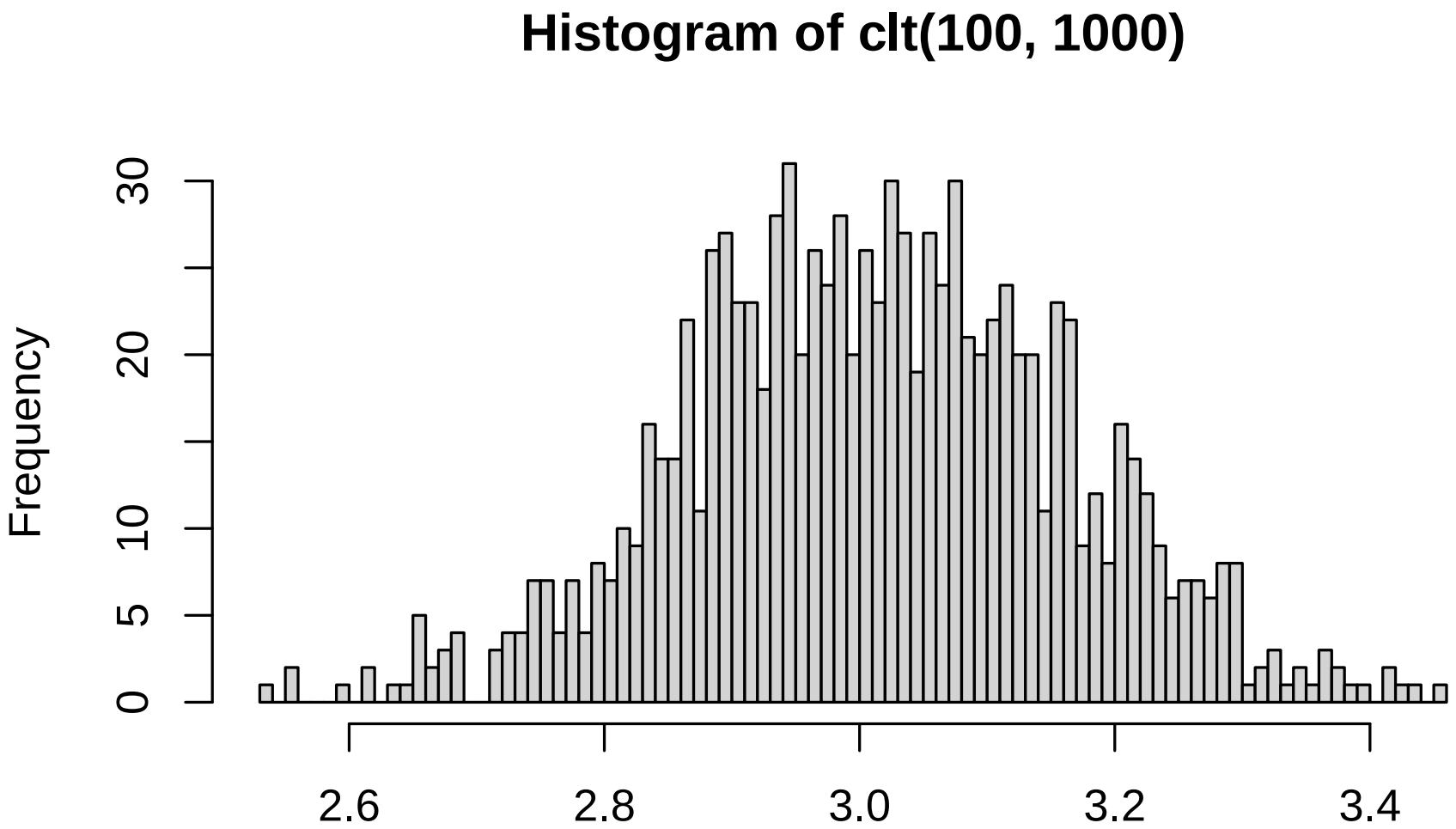
```
hist(clt(30,1000), breaks=100)
```



2.3. Central Limit Theorem (cont.)

100 observations: any difference from 30 observations?

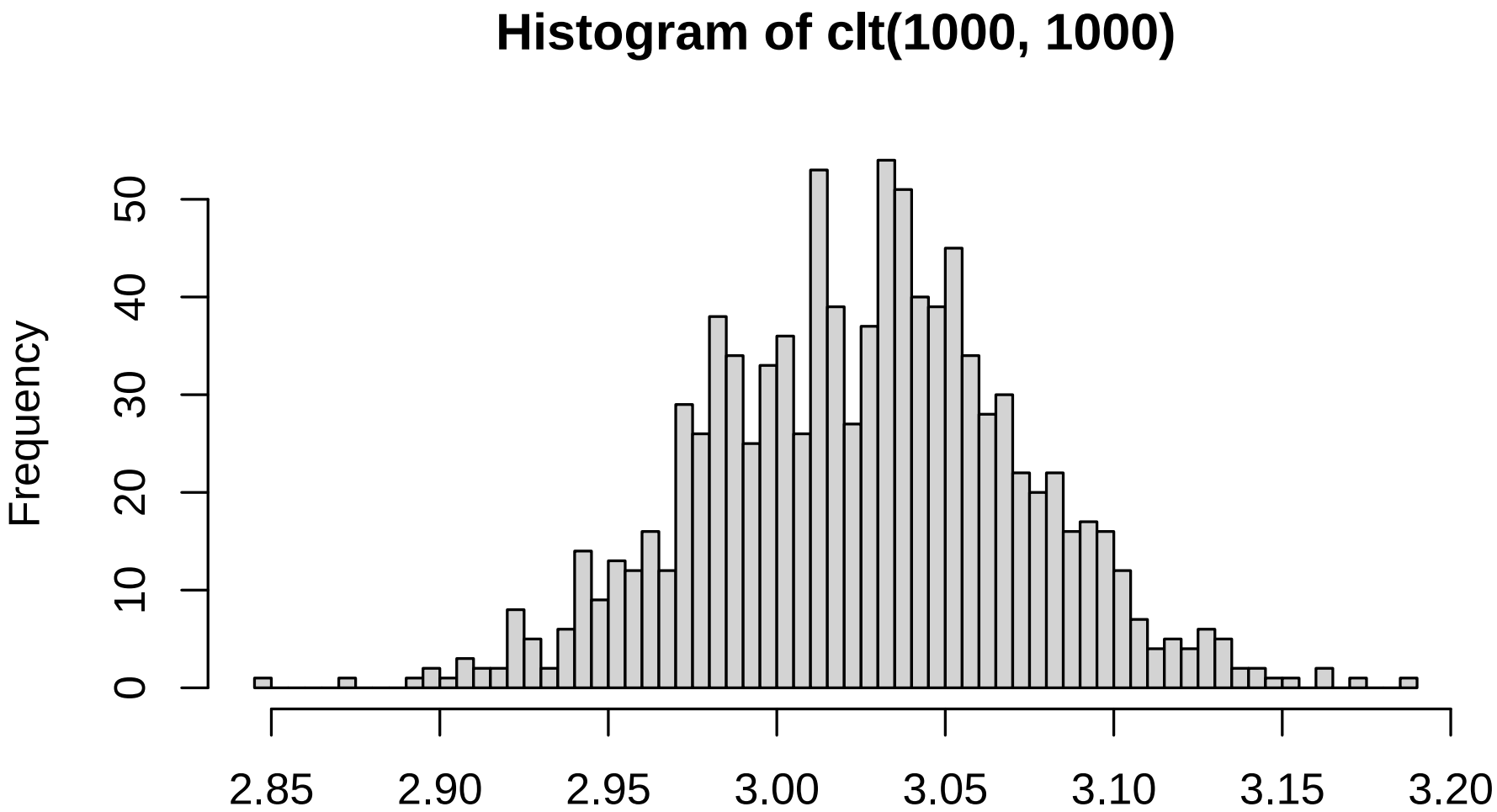
```
hist(clt(100,1000), breaks=100)
```



2.3. Central Limit Theorem (cont.)

1000 observations: how about now then?

```
hist(clt(1000,1000), breaks=100)
```



2.3. Central Limit Theorem (cont.)

Central Limit Theorem (CLT): The sample mean of any population with mean of μ and variance of σ follows $N(\mu, \frac{\sigma^2}{n})$ approximately when its sample size n approaches ∞. This distribution is called sampling distribution.

Also, an additional random variable $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, follows $N(0, 1)$ approximately when its sample size n approaches ∞.

We use the latter theorem more often when conducting a hypothesis testing.

2.3. Central Limit Theorem (cont.)

```
clt <- clt(1000,1000)
mean(clt)
```

```
## [1] 3.027237
```

```
sd(clt)^2
```

```
## [1] 0.002104577
```

```
sd(pop)^2/1000
```

```
## [1] 0.002135157
```

3. Statistical theory

Because we can estimate population mean and its variance only with our sample (better if our sample is large enough),

we know whether (and under what probability) the probability distribution function (PDF) of the sample mean contains a certain value.

1. If a certain value is zero, this is our null hypothesis ($H_0 : \bar{X} = 0$), and we are conducting one sample test.
2. If a certain value is another sample mean, this is our null hypothesis $H_0 : \bar{X_1} = \bar{X_2}$, and we are conducting two sample test.

We reject the null hypothesis if, under the null hypothesis, the probability of such a value as that which was actually observed (p-value) is less than or equal to a small, fixed pre-defined threshold value α, which is the level of significance.

3.1. Hypothesis Testing- One-Sample Test

```
sample <- sample(pop, 100)
mean(pop)
```

```
## [1] 3.0275
```

```
mean(sample)
```

```
## [1] 2.85
```

```
t.test(sample)
```

```
##
##  One Sample t-test
##
## data:  sample
## t = 17.908, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.534215 3.165785
## sample estimates:
## mean of x
##      2.85
```

3.2. Hypothesis Testing- Two-Sample Test

```r
sample1 <- sample(pop, 100)
sample2 <- sample(rbinom(10000, 10, 0.4), 100)

t.test(sample1, sample2)
```

```
##
##  Welch Two Sample t-test
##
## data:  sample1 and sample2
## t = -4.7169, df = 197.96, p-value = 4.515e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2904475 -0.5295525
## sample estimates:
## mean of x mean of y
##      2.82      3.73
```

3.3. A Pitfall of P-value

In various academic disciplines, p-value is now considered a great source of misreporting a finding of an empirical study:

https://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108

A few takeaways from the article:

1. By itself, p-value does not tell us anything about the size of the effect of interest. Researchers should bring as much contextual evidence as possible to buttress the claim.

2. A certain threshold (i.e. $p>0.05$) should not be taken for granted without any reference to the research design or the format of null hypothesis- the p-value can change even without any substantial change to the data or claims but only with those two above (so called p-hacking).

4. Linear Regression

The main reason for conducting linear regression is to find out the coefficient, which is basically a conditional mean of X given Y.

We test whether this conditional mean is equal to zero (one-sample, two-tailed test). This is the reason we look at p-value anyways!

But, as explained earlier, be careful of interpreting the p-value of coefficients.

4.1. Linear Regression in Scalar Form

Linear regression in scalar form is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + ... + \beta_k x_{ki} + \varepsilon_i$$

where $\epsilon \sim N(o, \sigma^2)$. This is our stochastic component. There should be no correlation between errors ($\mathbb{E}(\epsilon_i \times \epsilon_j) = 0$ for all $i \neq j$).

Meanwhile, our systematic component writes:

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + ... + \beta_k x_{ki}$$

4.2. Standard Error of the Regression

The variance of our error terms, σ^2, can be written as:

$$\begin{aligned}
\sigma^2 &= \mathbb{E}((\varepsilon_i - \mathbb{E}(\varepsilon_i))^2) \\
&= \mathbb{E}((\varepsilon_i - 0)^2) \\
&= \mathbb{E}(\varepsilon_i^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \\
&= \frac{1}{n} RSS
\end{aligned}$$

where RSS is the residual sum of squares. This is referred to as the standard error of the regression.

4.3. Square Root of the Mean of the Squared Errors (RMSE)

Also note that:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2}$$

which is the square root of the mean of the squared errors (RMSE). This is how much we expect our observation (y_i) to differ from its expected value ($\mathbb{E}(y_i)$), or our systematic component of the regression.

4.4. Linear Regression in Matrix Form

Linear regression in matrix form is:

$$Y = X\beta + \epsilon$$

which can be expanded as below:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

4.5. Vector of Stochastic Components

Note that we now have a vector of error terms, whose expectation is still zero:

$$\mathbb{E}(\epsilon) = \begin{bmatrix} \mathbb{E}(\varepsilon_1) \\ \mathbb{E}(\varepsilon_2) \\ \vdots \\ \mathbb{E}(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

4.6. Variance-Covarance Matrix (V-Cov)

Now I introduce (perhaps) the most important concept in matrix linear regression: variance-covariance matrix. It is a $n \times n$ matrix filled with variance and covariances of error terms.

$$\Sigma = \begin{bmatrix} \text{var}(\varepsilon_1) & \text{cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{cov}(\varepsilon_1, \varepsilon_n) \\ \text{cov}(\varepsilon_2, \varepsilon_1) & \text{var}(\varepsilon_2) & \cdots & \text{cov}(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\varepsilon_n, \varepsilon_1) & \text{cov}(\varepsilon_n, \varepsilon_2) & \cdots & \text{var}(\varepsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}(\varepsilon_1^2) & \mathbb{E}(\varepsilon_1 \varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_1 \varepsilon_n) \\ \mathbb{E}(\varepsilon_2 \varepsilon_1) & \mathbb{E}(\varepsilon_2^2) & \cdots & \mathbb{E}(\varepsilon_2 \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(\varepsilon_n \varepsilon_1) & \mathbb{E}(\varepsilon_n \varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_n^2) \end{bmatrix}$$

which can be reduced into:

$$\Sigma = \mathbb{E}(\epsilon \epsilon')$$

where ϵ' is a transpose of ϵ.

4.6. Variance-Covarance Matrix (V-Cov) (cont.)

Note that when the assumption $\mathbb{E}(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$ holds, Σ will look like:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

4.7. Beta Coefficient

We want to obtain the value for β which minimizes the residual sum of squares, hence this method being named after "least squares". Formally:

$$\arg\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = \arg\min_{\beta} \epsilon'\epsilon$$

Although we will not derive the whole process, it is known that β can be shortened as:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

This is our least squares estimator for true β.

4.7. Beta Coefficient (cont.)

Finally, we will conclude by emphasizing the importance of sample size in statistical analysis. But I will also show why the sample size is not a great concern for econometricians.

Assume that this is a true model for two independent variables, X, Z, and a dependent variable, Y.

$$Y = 1 + 0.5X + 0.3Z + \varepsilon$$

Where X and Z both follow a normal distribution:

```
x <- rnorm(10000, 5, 3)
z <- rnorm(10000, 3, 5)
e <- rnorm(10000, 5, 5)
y <- 1 + 0.5*x + 0.3*z + e
```

4.7. Beta Coefficient (cont.)

Assume that we had the n size of sample for each X, Z, and Y. It'd be great that we have larger sample, but let's compare each situation.

Now we are estimating a following model with each sample size we have:

$$Y = \alpha + \beta_1 X + \beta_2 Z + \varepsilon$$

Where $i = 1, 2, ..., n$, β_1 is a parameter estimate for the coefficient of X (which is 0.5 but we don't know), and β_2 is a parameter estimate for the coefficient of Z (which is 0.3 but again we don't know), and ε is an ideosyncratic error term.

Let's make a data frame before running a regression.

```
data <- data.frame(y=y, x=x, z=z)
```

4.7. Beta Coefficient (cont.)

1) When n = 10

```
sampledata <- dplyr::sample_n(data, 10)
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
ols <- lm(y ~ x + z, data=sampledata)
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ x + z, data = sampledata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8599 -3.0102  0.2035  2.0817  7.1176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.1866     2.4993   1.675    0.138
## x             0.6821     0.3640   1.874    0.103
## z             0.3813     0.4158   0.917    0.390
##
## Residual standard error: 4.46 on 7 degrees of freedom
## Multiple R-squared:  0.3449, Adjusted R-squared:  0.1577
## F-statistic: 1.843 on 2 and 7 DF,  p-value: 0.2276
```

4.7. Beta Coefficient (cont.)

2) When n = 30

```
sampledata <- dplyr::sample_n(data, 30)

ols <- lm(y ~ x + z, data=sampledata)
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ x + z, data = sampledata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3268  -3.9267   0.2261   3.1229  15.7339
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13685    2.14719   1.461  0.15558
## x            1.20274    0.36445   3.300  0.00272 **
## z            0.09635    0.25342   0.380  0.70676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.355 on 27 degrees of freedom
## Multiple R-squared:  0.2923, Adjusted R-squared:  0.2399
## F-statistic: 5.576 on 2 and 27 DF,  p-value: 0.009395
```

4.7. Beta Coefficient (cont.)

Again, don't get lured by the stars- remember our true parameters and see how these estimates are correct:

```
pe <- coefficients(ols)
vc <- vcov(ols)
coefs <- mvrnorm(10000, pe, vc)
mean(coefs[,2])
```

```
## [1] 1.212254
```

```
quantile(coefs[,2], c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.4861881 1.9466729
```

4.7. Beta Coefficient (cont.)

3) When n = 100

```
sampledata <- dplyr::sample_n(data, 100)

ols <- lm(y ~ x + z, data=sampledata)
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ x + z, data = sampledata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.024  -2.959  -0.697   3.532  11.610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.2710     1.0633   4.957 3.04e-06 ***
## x             0.5149     0.1744   2.953  0.00395 **
## z             0.4433     0.1035   4.282 4.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.379 on 97 degrees of freedom
## Multiple R-squared:  0.2125, Adjusted R-squared:  0.1963
## F-statistic: 13.09 on 2 and 97 DF,  p-value: 9.266e-06
```

4.7. Beta Coefficient (cont.)

Now it gets much closer to the population parameters. But consider the next example:

4) When n = 100, but without z

```
sampledata <- dplyr::sample_n(data, 100)
ols <- lm(y ~ x, data=sampledata)
summary(ols)
```

```
##
## Call:
## lm(formula = y ~ x, data = sampledata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4404  -4.6614   0.1052   4.7292  13.5235
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.7211     1.0749   6.253 1.06e-08 ***
## x             0.5537     0.1908   2.902  0.00458 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.925 on 98 degrees of freedom
## Multiple R-squared:  0.07913,    Adjusted R-squared:  0.06973
## F-statistic: 8.421 on 1 and 98 DF,  p-value: 0.004581
```