BUSN 5000

Making Inferences

Chris Cornwell

Terry College of Business

Fall 2023

Section 1

Preliminaries

Week 4 :: Reading

Reading: Bekes and Kedzi, chs 5-6; Bueno de Mesquita and Fowler, ch 6

IRL: "Building a Culture of Experimentation", *Harvard Business Review*, March-April 2021.

R highlights

No new packages this week.

Functions

- `qnorm` for quartiles of the standard normal distribution
- `geom_pointrange` for plotting vertical intervals with point estimates
- `t.test` for t tests

The main things

1. Frequentism
2. Consistency
3. Sampling distribution
4. Standard error
5. Asymptotic normality
6. Confidence interval
7. t test
8. p-value

A/B testing at Booking.com

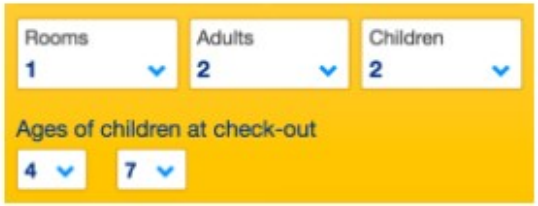
Scenario #2

Hypothesis

Displaying the checkout date when users select the age of children in their party improves their experience.

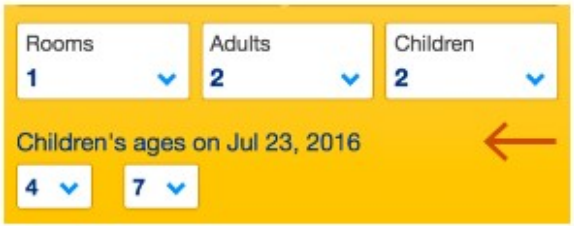
A: The Control

Shows the site's current practice



B: The Treatment

Adds the checkout date above children's ages



The Result

The treatment had a significant positive impact on the key metric, and the change is implemented.

Section 2

From earnings to wages

What do we want to learn?

We want to shift the focus from earnings to wages. Why? Because the story is really about productivity. Earnings involves both wages and hours and therefore masks the underlying productivity story.

We want to learn about the disparity in wages and wage profiles and how they vary with education, just as we did for earnings.

We also want learn about the role of hours. Is the gender gap related to how many hours you work? Do you have a model for how this relationship might go?

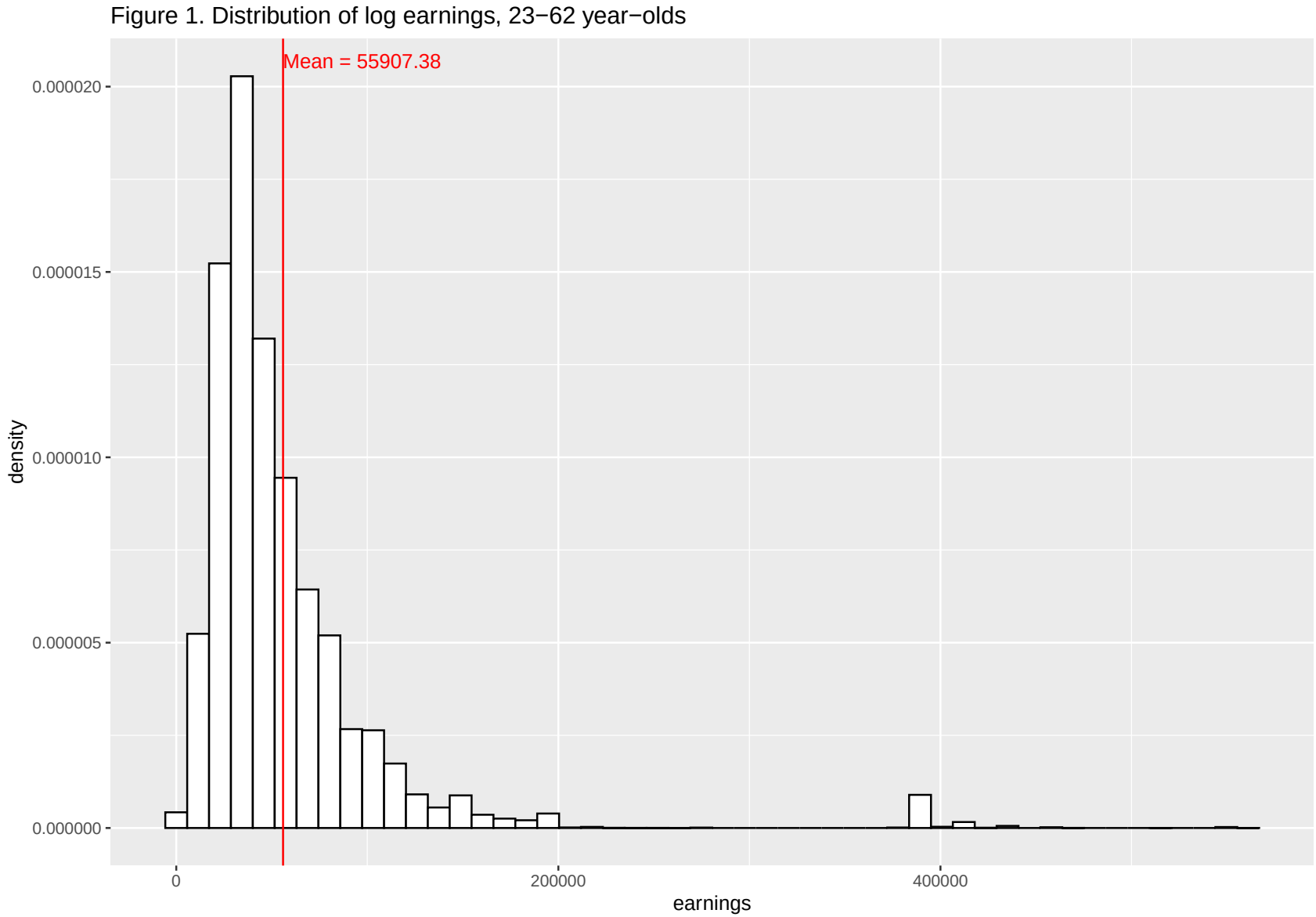
Facts from the 2009 March CPS analysis sample

```
cps_mar <- read_xlsx("data/cps09mar.xlsx")
cps_mar_2362 <- cps_mar %>%
  filter(age >= 23,
         age <= 62)
datasummary(
  earnings + age ~ N + Mean + SD,
  data=cps_mar_2362,
  title="Earnings and age summary statistics, 23-62 year-olds")
```

Table 1: Earnings and age summary statistics, 23-62 year-olds

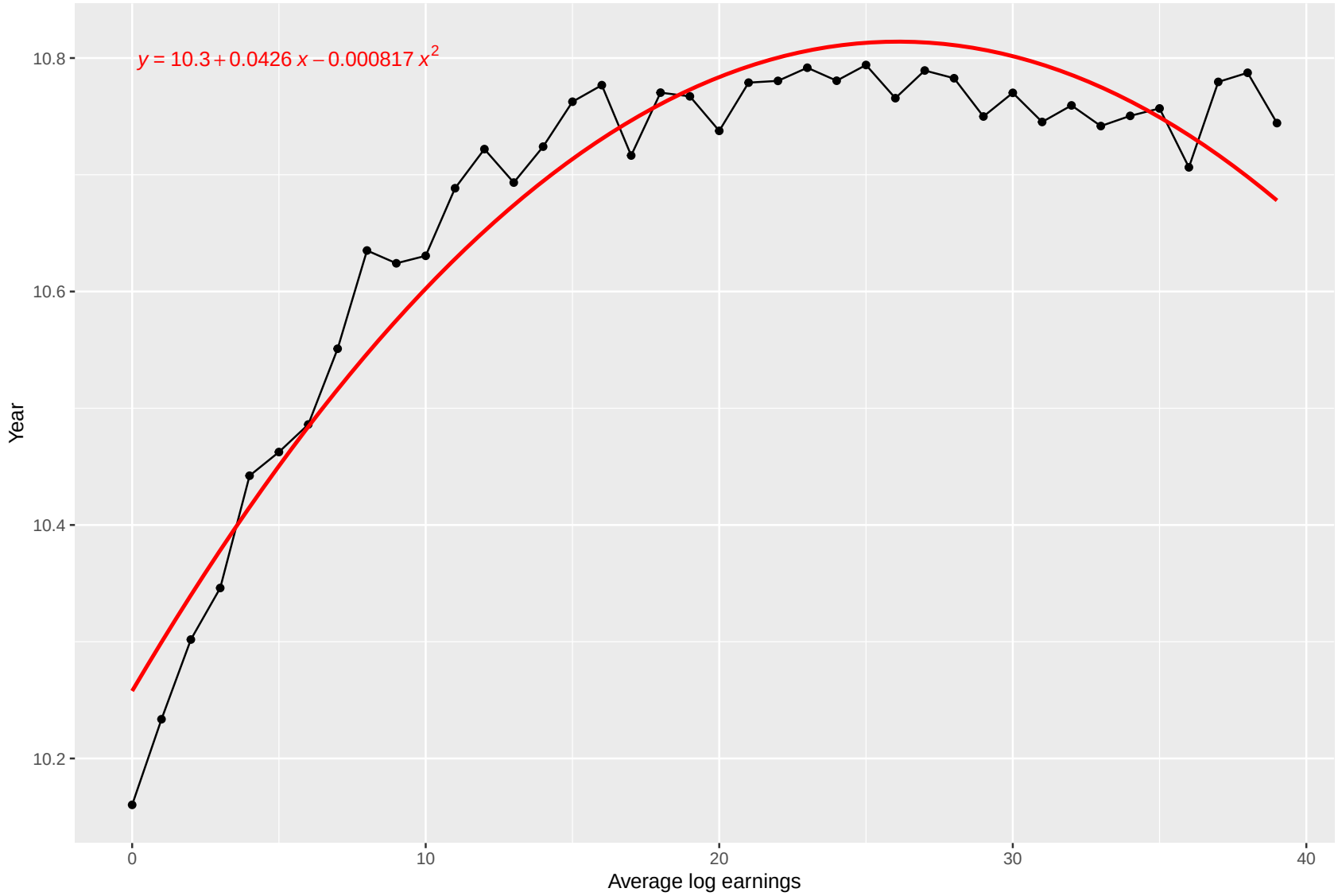
	N	Mean	SD
earnings	47671	55 907.38	52 270.82
age	47671	41.96	10.27

Log earnings distribution, March 2009 CPS



$\hat{E}(learnings|age)$ using March 2009 analysis sample

Figure 2. Log earnings profile



From earnings to wages

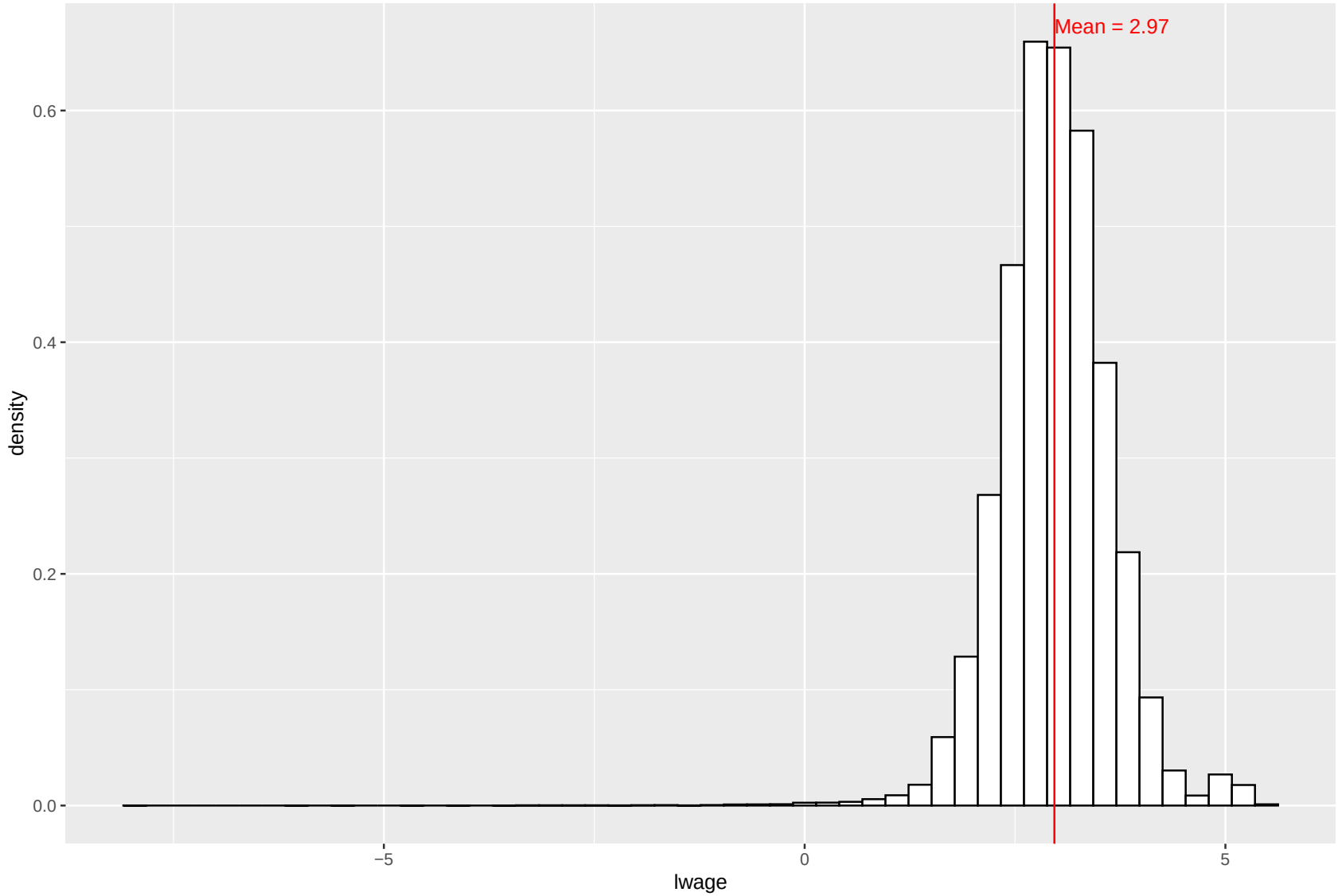
```
cps_mar_2362 <- cps_mar_2362 %>%
  filter(earnings>0) %>%
  mutate(wage   = earnings/(hours*week),
    lwage = log(wage),
    gender = ifelse(female == 1, "Female", "Male"),
    over40 = ifelse(hours>40, ">40", "<=40"))
datasummary(earnings + wage + lwage + age + hours ~ gender*(N + Mean + SD),
  data=cps_mar_2362,
  title="Summary statistics by gender, 23-62 year-olds")
```

Table 2: Summary statistics by gender, 23-62 year-olds

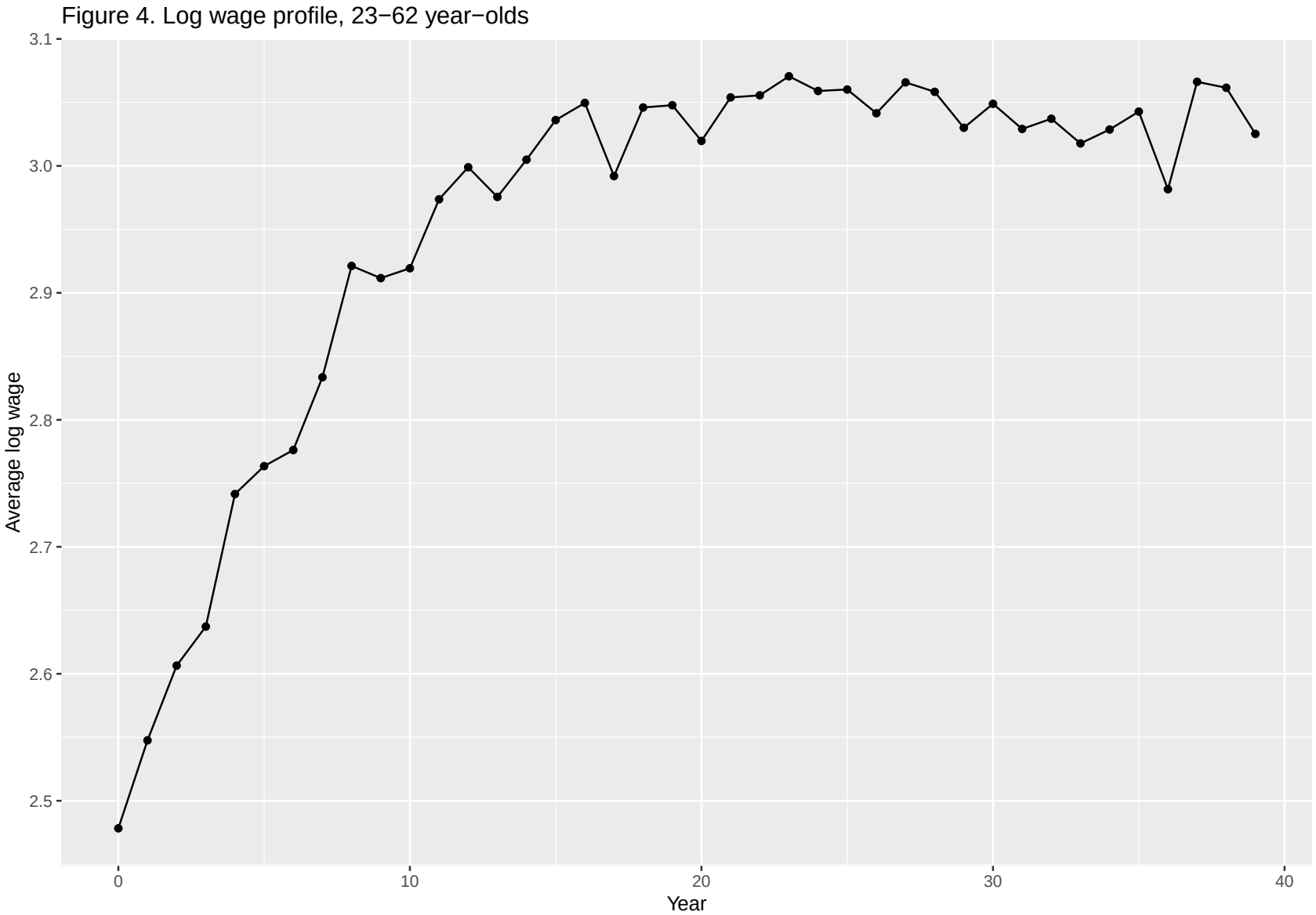
	Female			Male		
	N	Mean	SD	N	Mean	SD
earnings	20392	44 827.78	36 585.22	27279	64 189.77	60 115.22
wage	20392	20.26	15.25	27279	27.21	23.53
lwage	20392	2.83	0.60	27279	3.07	0.68
age	20392	42.11	10.35	27279	41.86	10.20
hours	20392	42.38	6.24	27279	45.02	8.53

Log wage distribution, March 2009 CPS

Figure 3. Distribution of log wages, 23−62 year−olds



$\hat{E}(lwage|age)$ using March 2009 analysis sample



Section 3

Evaluating Estimates

Frequentist framework

Again, the point of all of this is learning about the world using data. Our framework is gathering *random samples* from the *population* of interest to infer features of the data-generating process.

By "features", we mean at least expected value and variance. When there are relationships among random variables to sort out (say, between a Y and some Xs), there are even more things to learn about.

Every feature is, in principle, an *estimand* we learn about by applying *estimators* we calculate using data we've collected. The calculations produce *estimates* that can inform us about the estimand – if the estimators are *good*.

Good estimators

Remember, in general, we can think of all estimators like this:

$$estimator = estimand + bias + sampling\ error$$

We're going to say that an estimator is good if two things are true:

1. The estimator approaches the estimand as sample size increases.
2. We can treat the estimator's sampling distribution as normal for large sample sizes.

We really need these two things to make the inferences our analyses call for.

Consistency

The first thing is called *consistency* and, roughly speaking, holds if the LLN applies. More practically, it means that bias and sampling error approach zero as sample size increases.

For example, consider $\overline{Y}$, the usual estimator of $\mu = E(Y)$. We can say that $\overline{Y}_N$ is consistent if

$$\overline{Y}_N \xrightarrow{p} \mu \quad \text{as } N \to \infty,$$

where the p over the arrow indicates "convergence in probability". The LLN says this will be true under random sampling from a population with mean, μ.

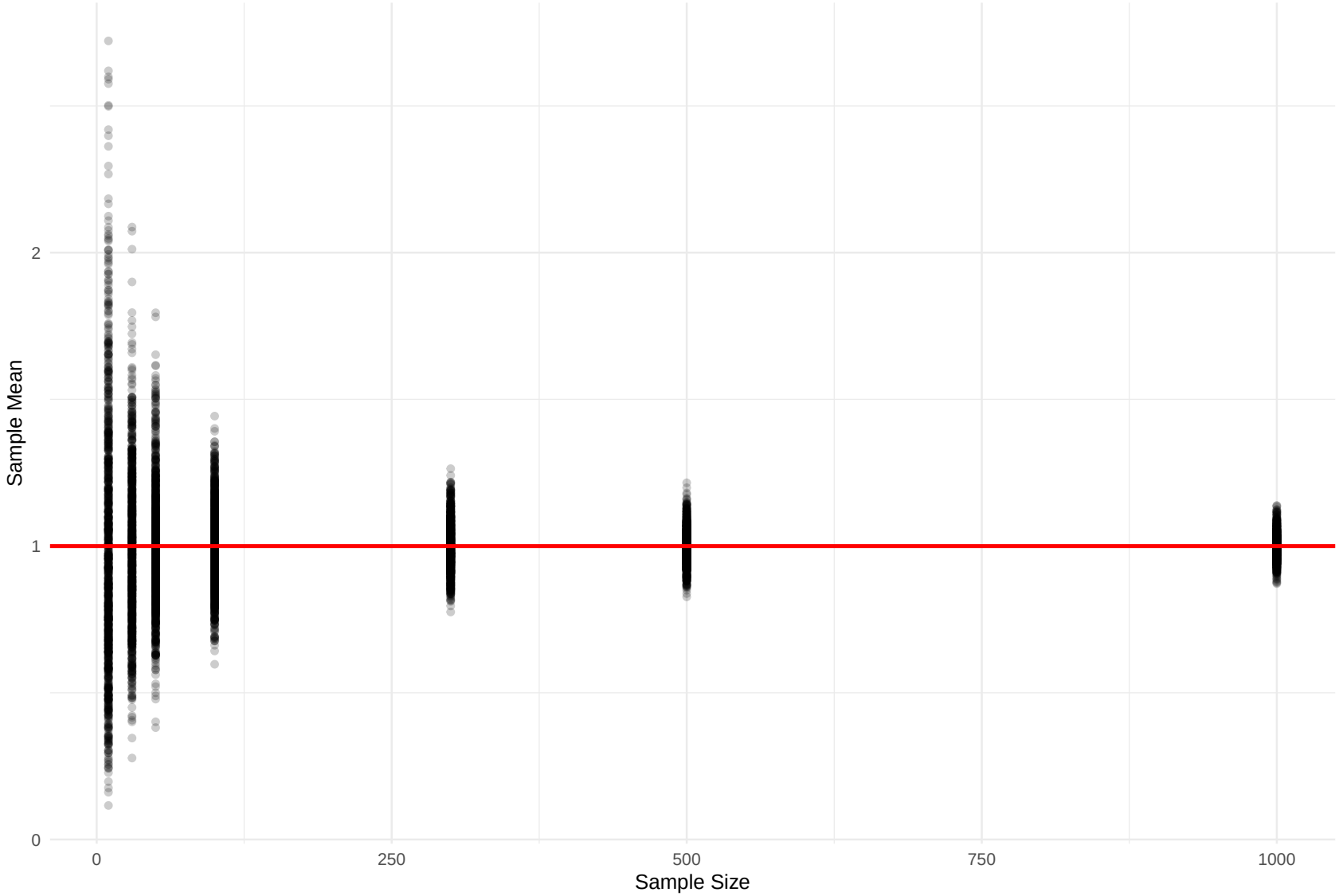
Simulating the LLN

```r
# Initialize parameters
n_trials <- 1000 # Number of Monte Carlo trials for each sample size
sample_sizes <- c(10, 30, 50, 100, 300, 500, 1000) # Different sample sizes to consider
# true_mean <- 0 # True mean of the normal distribution
true_mean <- 1 # True mean of the chi-sq distribution
# Initialize an empty data frame to store results
results <- data.frame(sample_size = integer(), sample_mean = numeric())
# Perform Monte Carlo simulation
for (n in sample_sizes) {
  for (i in 1:n_trials) {
#    sample_data <- rnorm(n, mean = true_mean, sd = 1) # Generate a sample
    sample_data <- rchisq(n, 1)  # Generate a sample from Chi-square distribution
    sample_mean <- mean(sample_data) # Calculate sample mean

    # Append to results data frame
    results <- rbind(results, data.frame(sample_size = n, sample_mean = sample_mean))
  }
}
# Plot the results
lln <- ggplot(results, aes(x = sample_size, y = sample_mean)) +
  geom_point(alpha = 0.2) +
  geom_hline(yintercept = true_mean, color = "red", size = 1) +
  ggtitle("Figure 5. Monte Carlo Demonstration of the Law of Large Numbers") +
  xlab("Sample Size") +
  ylab("Sample Mean") +
  theme_minimal()
```

Visualizing the simulation

Figure 5. Monte Carlo Demonstration of the Law of Large Numbers



Sampling distributions

Consistency is an essential estimator property, but it tells us nothing about an estimator's *sampling distribution*, and we can't really evaluate our estimates, let alone do *hypothesis testing*, without it.

Think of the sampling distribution as the distribution you would get if you computed the estimator an infinite number of times. It describes the likelihood the estimator takes on different values across different random samples.

The problem is we rarely know it, so we must rely on approximations. This is where the *Central Limit Theorem (CLT)* comes in.

The CLT to the rescue

Let's stay with $\overline{Y}_N$.

The CLT says that under random sampling from population with mean μ and variance σ^2, a *standardized* $\overline{Y}_N$ (call it Z_N) converges to a random variable (Z) with a standard normal distribution:

$$Z_N = \frac{\overline{Y}_N - \mu}{\sqrt{\text{var}(\overline{Y}_N)}} = \frac{\overline{Y}_N - \mu}{\sqrt{\frac{\sigma^2}{N}}} = \frac{\sqrt{N}(\overline{Y}_N - \mu)}{\sigma} \xrightarrow{d} Z, \quad \text{as } N \to \infty,$$

The factor $\sqrt{N}$ is important because the distribution of $\frac{(\overline{Y}_N - \mu)}{\sigma}$ literally collapses to zero as $N \to \infty$. Multiplying by $\sqrt{N}$ guarantees that the variance of Z_N remains constant because $(\overline{Y}_N - \mu) \to 0$ at the same rate $\sqrt{N} \to \infty$. This way we can learn something about the distribution of Z_N as sample size grows.

Asymptotic normality

The practical value of the CLT is in *approximating* the unknown distribution of $\overline{Y}_N$:

$$Z_N \to N(0,1) \quad \Rightarrow \quad \overline{Y}_N \overset{A}{\sim} N\left(\mu, \frac{\sigma^2}{N}\right),$$

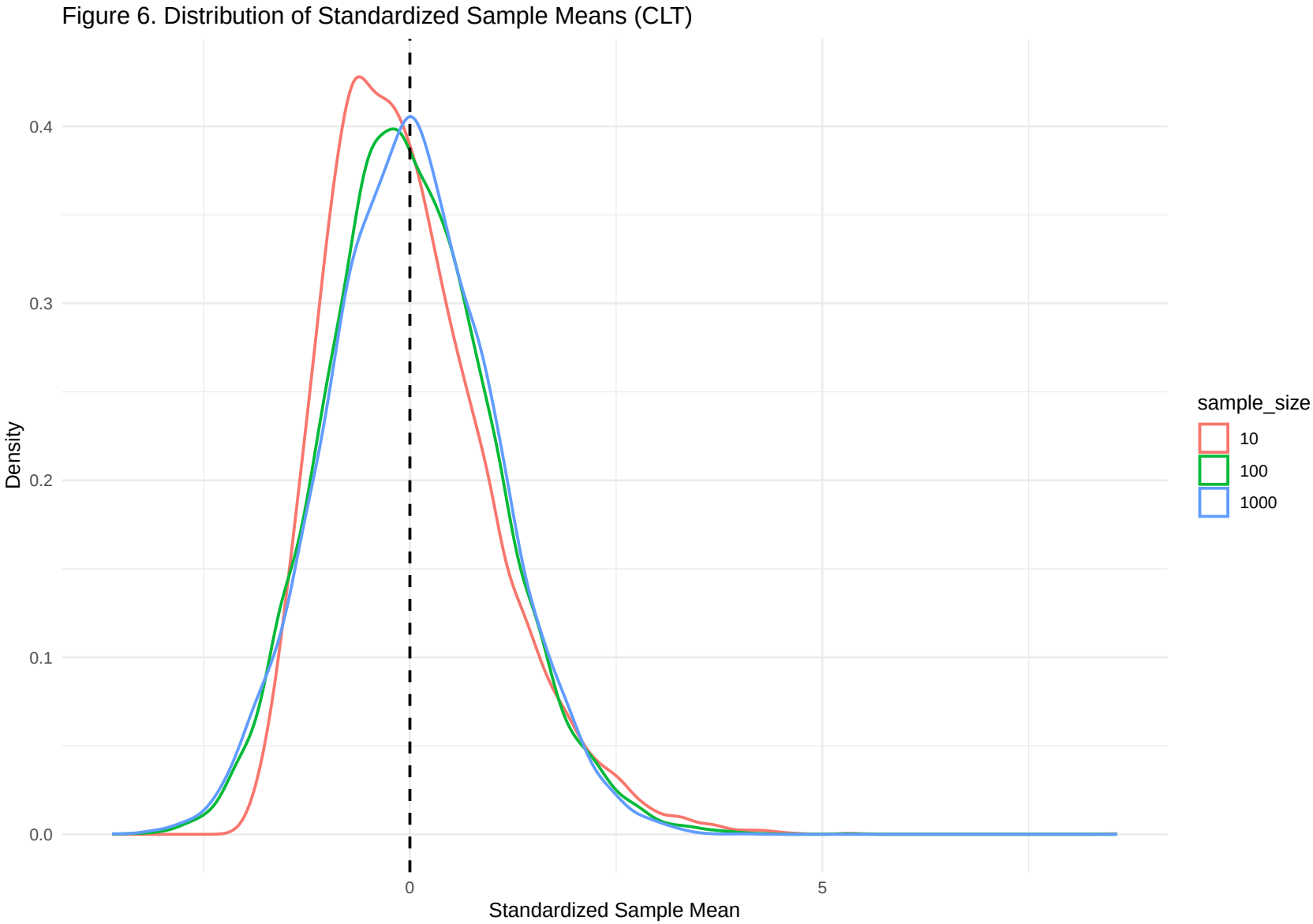
where the A over the $\sim$ can be read as "approximately" or "asymptotically".

We say that $\overline{Y}$ is approximately or asymptotically normally distributed or has the property of *asymptotic normality*.

Simulating the CLT

```r
# Initialize parameters
n_trials <- 10000  # Number of Monte Carlo trials for each sample size
sample_sizes <- c(10, 100, 1000)  # Different sample sizes to consider
true_mean <- 1  # True mean of the Chi-square distribution with 1 degree of freedom
true_var <- 2  # True variance of the Chi-square distribution with 1 degree of freedom
# Initialize an empty data frame to store results
result_df <- data.frame(standardized_mean = numeric(),
                        sample_size = factor())
# Perform Monte Carlo simulation and collect sample means
for (n in sample_sizes) {
  standardized_means <- numeric(n_trials)  # Vector to store standardized sample means
  for (i in 1:n_trials) {
    sample_data <- rchisq(n, 1)  # Generate a sample from Chi-square distribution
    sample_mean <- mean(sample_data)  # Calculate sample mean
    sample_sd <- sqrt(var(sample_data)) # Calculate sample standard deviation
    standardized_means[i] <- sqrt(n) * (sample_mean - true_mean) / sqrt(true_var)
  }
  temp_df <- data.frame(standardized_mean = standardized_means,
                        sample_size = factor(rep(n, n_trials)))
  result_df <- rbind(result_df, temp_df)
}
# Create density plot of standardized sample means
p <- ggplot(result_df, aes(x = standardized_mean, color = sample_size)) +
  geom_density(size = .75) +
  geom_vline(aes(xintercept = 0), color = "black", linetype = "dashed", size = .75) +
  ggtitle("Figure 6. Distribution of Standardized Sample Means (CLT)") +
  xlab("Standardized Sample Mean") +
  ylab("Density") +
  theme_minimal()
```

Visualizing the simulation



Estimating var($\overline{Y}$)

The CLT holds even if var($\overline{Y}$) is replaced by a consistent estimator. The standard choice is to replace σ^2 with

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2.$$

and estimate var($\overline{Y}$)

$$\widehat{\text{var}}(\overline{Y}) = \frac{s^2}{N}.$$

Using $\hat{\sigma}^2$ would be fine too because it is consistent, which is all we need for *asymptotically valid* inference.

Either way, we call $\widehat{\text{var}}(\overline{Y})$ the *estimated variance* of var($\overline{Y}$), its square root, $\sqrt{\widehat{\text{var}}(\overline{Y})}$ the estimated *standard error*.

Confidence interval for the sample mean

A *confidence interval (CI)* tells you how likely an estimate is close to its target in the population.

Using the CLT, we can say things like this about the sample mean:

$$P\left(-1.96 < \frac{\overline{Y} - \mu}{s/\sqrt{N}} < 1.96\right) = .95,$$

where 1.96 is the 2.5% critical value for the standard normal distribution (see Figure 7 on the next slide). This implies the 95% CI for the sample mean can be constructed as

$$\overline{Y} \pm 1.96 \frac{s}{\sqrt{N}}.$$

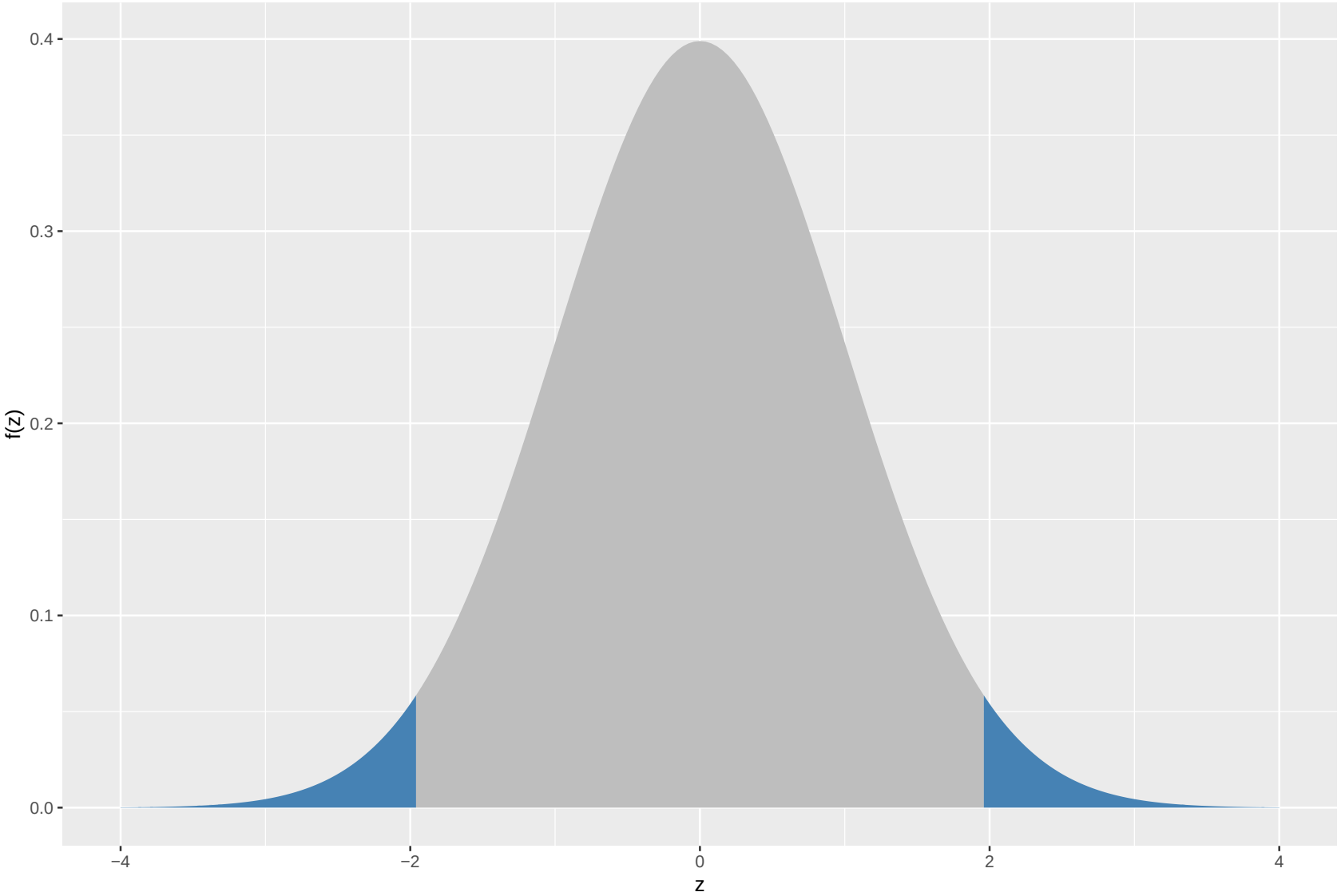
For a given $\overline{Y}$ (an estimate based on a particular sample), this *interval estimator* produces an *interval estimate* for μ. The interpretation of 95% CI interval is a random interval that will contain the true value of μ in 95% of samples.

Standard normal critical values

```
c_h <- qnorm(.975)
c_l <- qnorm(.025)
ggplot(data.frame(x=c(-4,4)), aes(x)) +
  stat_function(fun = dnorm,
                geom="area",
                fill = "steelblue",
                xlim = c(-4, c_l)) +
  stat_function(fun = dnorm,
                geom="area",
                fill = "grey",
                xlim = c(c_l, c_h)) +
  stat_function(fun = dnorm,
                geom="area",
                fill = "steelblue",
                xlim = c(c_h, 4)) +
  xlab("z") +
  ylab("f(z)") +
  labs(title="Figure 7. 2.5% and 97.5% critical values for N(0,1)",
    x="z",
    y="f(z)")
```

Standard normal critical values visualized

Figure 7. 2.5% and 97.5% critical values for the standard normal distribution



Confidence intervals in general

In general, if we have an estimator of some population estimand and the CLT applies, we can proceed as if the distribution of the estimator is normal:

$$estimator \overset{A}{\sim} N[estimand, \text{var}(estimator)],$$

which allows us to say

$$\frac{estimator - estimand}{\hat{se}(estimator)} \sim N(0, 1)$$

in large samples, where $\hat{se}(estimator) = \sqrt{\widehat{\text{var}}(estimator)}$ is the estimated *standard error* of $\hat{\theta}$.

This justifies statements like

$$P\left(-1.96 < \frac{estimator - estimand}{\hat{se}(estimator)} < 1.96\right) = .95,$$

leading to CIs like this

$$estimator \pm 1.96\, \hat{se}(estimator).$$

Hypothesis testing

Statistical hypothesis tests are basically just CIs viewed from another angle. Hypothesis testing translates the information contained in a CI into yes/no answers to particular questions like, "Are average wages the same for men and women?".

As you may remember, statistical tests involve two hypotheses: the null (H_0) and the alternative (H_1). The null is the hypothesis of interest, as it specifies what is being tested.

When we perform hypothesis tests, we can go wrong in two ways:

1. Reject a true null or commit a *Type-I error* ("false positive").
2. Fail to reject a false null or commit a *Type-II error* ("false negative").

Sometimes the courtroom analogy can sort these out. There, H_0 is that the defendant is innocent. It is the burden of the prosecution to provide evidence sufficient to reject it. A *Type-I* error occurs when an innocent person is convicted. Allowing a guilty person to go free is a *Type-II* error.

Asymptotically valid test statistics

Hopefully we have made it clear that we depend on *asymptotic analysis* for statistical inference. An asymptotically valid test statistic is easily derived from the fact that, by the CLT,

$$\frac{estimator - estimand}{\hat{se}(estimator)} \sim N(0,1).$$

To test the null that the estimand equals some hypothesized value, the *test statistic* would be

$$\frac{estimate - null\ value}{\hat{se}(estimator)} \sim N(0,1).$$

For a two-tailed test, you compare the value of the test statistic with the appropriate *critical value* from the normal distribution (like 1.96 for a 5% significance level test). If the (absolute) value of the test-statistic is greater than critical value, the null is rejected.

t test for population mean

You may recognize what we just described as a t test because it is. The test statistic is labeled "t" because when your random sample comes from a normal population, it has a t distribution.

However, we won't make that assumption, and besides, the t distribution becomes indistinguishable from the normal for large N.

In any event, let's translate to the case of the sample mean. The estimand is μ and the estimator $\overline{Y}$, whose standard error is $\frac{s}{\sqrt{N}}$. So the t statistic for, say, $H_0 : \mu = \mu_0$ is

$$t = \frac{\overline{Y} - \mu_0}{s/\sqrt{N}}.$$

We would then reject the null at the 5% level if $|t| > 1.96$.

Section 4

Evaluating the CEF

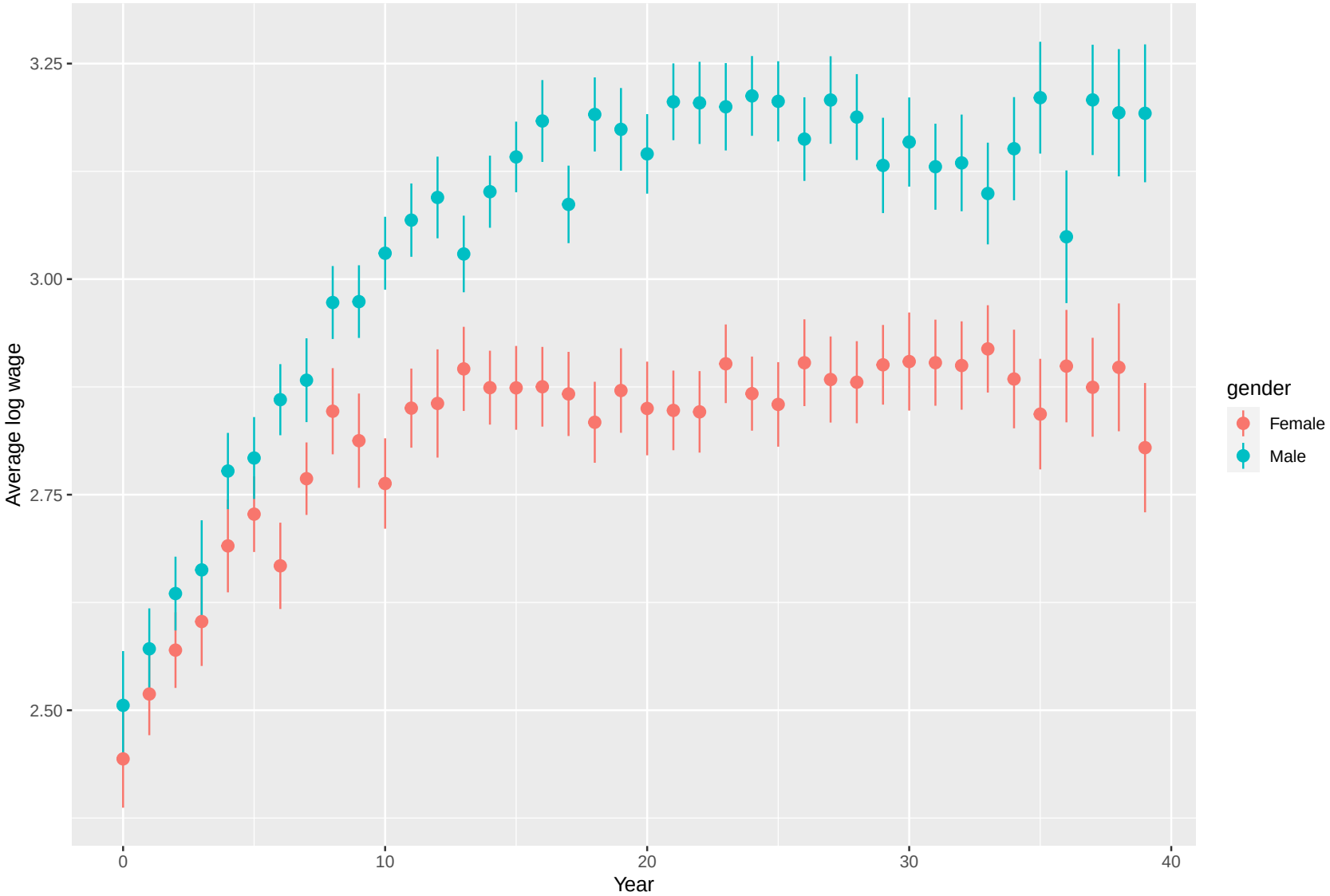
Confidence intervals for the CEF

Now, let's use our understanding of sampling distributions, standard errors and confidence intervals to show the statistical uncertainty about our CEF *point estimates*. Does the male conditional mean estimate's CI ever contain the female estimate?

```
cef_fvm_w <- cps_mar_2362 %>%
  mutate(age=age-23) %>%
  group_by(age, gender) %>%
  summarise(
    lwage_bar = mean(lwage, na.rm = TRUE),
    lwage_se   = sd(lwage, na.rm = TRUE)/sqrt(n()),
    upper = lwage_bar + lwage_se*1.96,
    lower = lwage_bar - lwage_se*1.96
    )
cef_fvm_w_p <- ggplot(cef_fvm_w, aes(age, lwage_bar, color=gender)) +
  geom_pointrange(aes(ymin = lower, ymax = upper)) +
  labs(
    title="Figure 8. Log wage profiles by gender, 23-62 year-olds",
    x="Year",
    y="Average log wage"
    )
```

Log wage CEFs by gender

Figure 8. Log wage profiles by gender, 23−62 year−olds



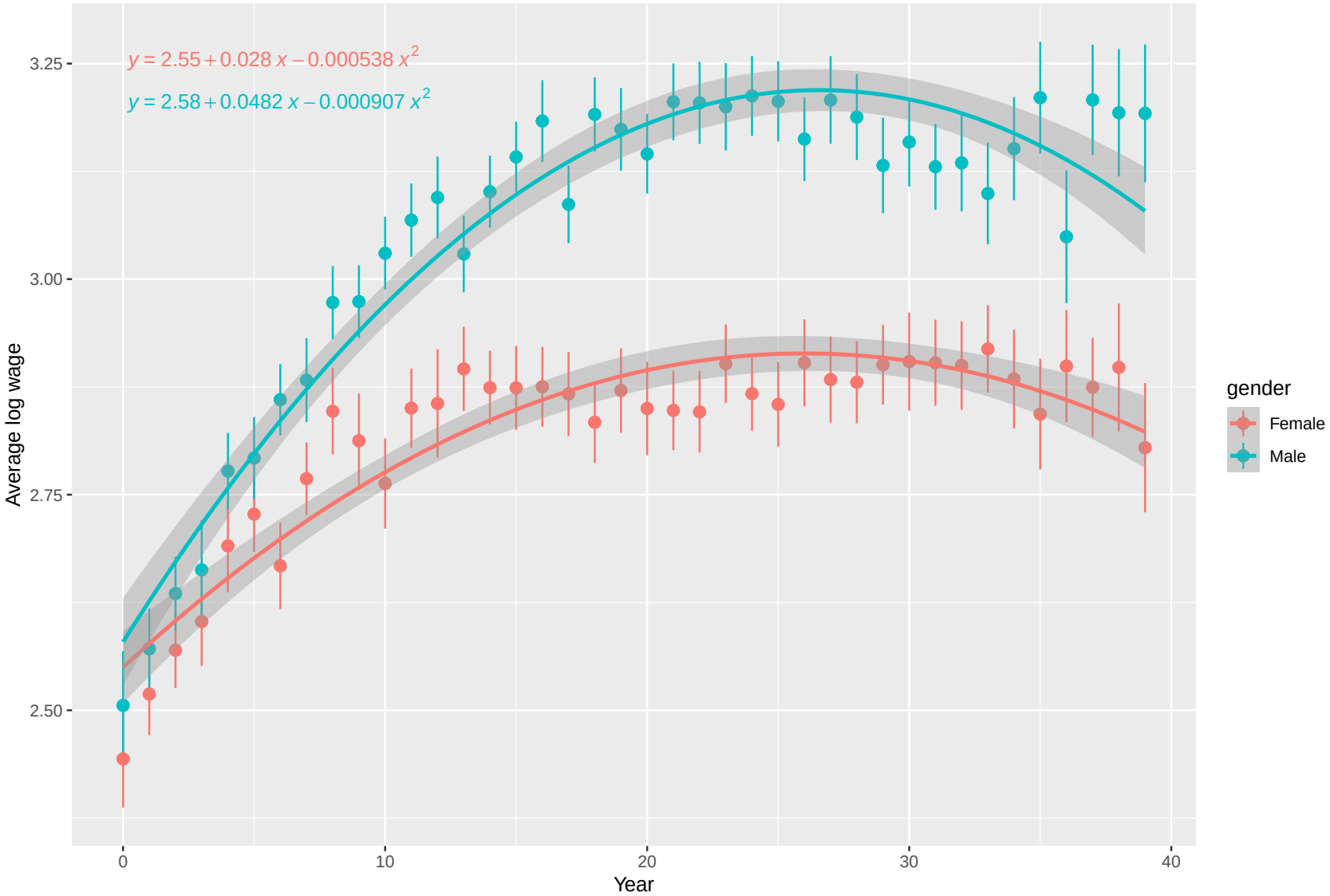
Quadratic fit with standard errors

Next, we'll add a quadratic fit for $E(log\ wage|age)$ with standard error bands for fitted model. We will have more to say about this in Part II.

```
formula <- y~poly(x,2, raw=TRUE)
cef_fvm_w_fit <- cef_fvm_w_p +
  geom_smooth(
    method = "lm",
    formula = formula,
    aes(group = gender),
    se = TRUE
    ) +
  stat_poly_eq(
    aes(label =  paste(stat(eq.label))),
    formula = formula,
    parse = TRUE
    )
```

Log wage CEF quadratic fits by gender

Figure 8. Log wage profiles by gender, 23−62 year−olds



Is the estimated gender gap statistically significant?

Stated as a null hypothesis, we want to test

$$H_0 : \mu_{men} - \mu_{women} = 0,$$

where $\mu_g = E(log\ wage|gender)$, $g = men,\ women$.

The test statistic for H_0 is

$$t = \frac{\hat{\mu}_{men} - \hat{\mu}_{women} - 0}{\hat{se}(\hat{\mu}_{men} - \hat{\mu}_{women})} = \frac{\hat{\mu}_{men} - \hat{\mu}_{women} - 0}{\sqrt{\frac{s^2_{men}}{N_{men}} + \frac{s^2_{women}}{N_{women}}}}.$$

Conducting a t test with `t.test`

The test statistic for the null that there is no difference in average log wages between men and women can be easily computed with `t.test` function. There shouldn't be any surprise about how this turns out.

```
##
##  Welch Two Sample t-test
##
## data:  lwage by gender
## t = -40.948, df = 46542, p-value < 0.0000000000000022
## alternative hypothesis: true difference in means between group Female an
## 95 percent confidence interval:
##  -0.2527856 -0.2296916
## sample estimates:
## mean in group Female   mean in group Male
##             2.829732             3.070971
```

p values

Instead of setting the significance level in advance – say, for example, at 5% – and letting it drive decision-making, we can report p values. The p value, or *marginal significance level*, is the probability of drawing a test statistic at least as adverse to the null as the one you actually calculated, conditional on the null being true:

$$p = P(\text{test-stat} \geq \text{test-stat value} | H_0).$$

Put differently, the p value is the largest significance level at which I could conduct the test and still fail to reject the null.

Just as larger values of the test statistic are associated with greater evidence against the null, so are smaller p values.

Statistical vs practical significance

Reporting p values can bring clarity to the discussion of statistical and practical significance.

A large test statistic value speaks to statistical significance, while practical significance is tied up in the magnitude of the parameter estimate.

A t statistic can be large either because the difference between the estimate and null value is large or its estimated standard error is small. It will be wise to take both into consideration when interpreting the results of statistical inference.

Depending on sample size, a p value larger than .05 may not lead to accepting the null.

Section 5

Hours of work and the gender gap

Wages and hours

Let's establish some basic empirical facts about our sample of 23-62 year-olds, highlight hours of work.

Table 3: Summary statistics by under/over 40 hours per week

	<=40			>40		
	N	Mean	SD	N	Mean	SD
wage	32313	22.16	17.53	15358	28.59	25.59
lwage	32313	2.91	0.61	15358	3.09	0.74
age	32313	41.68	10.44	15358	42.56	9.87
hours	32313	39.83	0.71	15358	52.43	8.81
female	32313	0.49	0.50	15358	0.31	0.46

CEFs by under/over 40 hours per week and gender

Figure 9. Log wage profiles by gender, under/over 40 hours per week

Documenting the gender gap among 23-30 year-olds who work >40 hours

```
cps_mar_2330 <- cps_mar_2362 %>%
  filter(
    over40==">40",
    age>=23 & age<=30
    )
datasummary(
  wage + lwage ~ gender*(N + Mean + SD),
  data=cps_mar_2330,
  title="Summary statistics, 23-30 year-olds working over 40 hours"
  )
t.test(lwage~gender, data=cps_mar_2330)
```

Testing the null of no gender gap

Table 4: Summary statistics, 23-30 year-olds working over 40 hours

	Female			Male		
	N	Mean	SD	N	Mean	SD
wage	638	17.07	9.15	1490	19.36	13.25
lwage	638	2.69	0.59	1490	2.79	0.66

```
##
##  Welch Two Sample t-test
##
## data:  lwage by gender
## t = -3.3861, df = 1347.3, p-value = 0.0007293
## alternative hypothesis: true difference in means between group Female an
## 95 percent confidence interval:
##  -0.15448031 -0.04114576
## sample estimates:
## mean in group Female   mean in group Male
##             2.689468             2.787281
```