Lesson 1: Review

Nicky Wakim

2025-01-06

What did we learn in 511/611? (1/2)

- In 511, we talked about *categorical* and *continuous* outcomes (dependent variables)

- We also talked about their relationship with 1-2 *continuous* or *categorical* exposure (independent variables or predictor)

- We had many good ways to assess the relationship between an outcome and exposure:

	Continuous Outcome	Categorical Outcome
Continuous Exposure	Correlation, simple linear regression	??
Categorical Exposure	t-tests, paired t-tests, 2 sample t-tests, ANOVA	proportion t-test, Chi-squared goodness of fit test, Fisher's Exact test, Chi-squared test of independence, etc.

What did we learn in 511/611? (2/2)

- You set up a really **important foundation**
  - Including distributions, mathematical definitions, hypothesis testing, and more!

- Tests and statistical approaches learned are incredibly helpful!

- While you had to learn a lot of different tests and approaches for each combination of categorical/continuous exposure with categorical/continuous outcome
  - **Those tests cannot handle more complicated data**

- **What happens when other variables influence the relationship between your exposure and outcome?**
  - Do we just ignore them?

What will we learn in this class?

- We will be building towards models that can handle many variables!

  - **Regression** is the building block for modeling multivariable relationships

- In Linear Models we will *build*, *interpret*, *and evaluate* **linear regression models**

Process of regression data analysis

Model Selection

- Building a model
- Selecting variables
- Prediction vs interpretation
- Comparing potential models

Model Fitting

- Find best fit line
- Using OLS in this class
- Parameter estimation
- Categorical covariates
- Interactions

Model Evaluation

- Evaluation of model fit
- Testing model assumptions
- Residuals
- Transformations
- Influential points
- Multicollinearity

Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X

Main sections of the course

1. Review
2. Simple Linear Regression
    - Model evaluation and Model use
3. Intro to MLR: estimation and testing
    - Model use
4. Diving into our predictors: categorical variables, interactions between variable
    - Model fitting
5. Key ingredients: model evaluation, diagnostics, selection, and building
    - Model evaluation and Model selection

Main sections of the course

1. Review

2. Intro to SLR: estimation and testing
   - Model fitting

3. Intro to MLR: estimation and testing
   - Model fitting

4. Diving into our predictors: categorical variables, interactions between variable
   - Model fitting

5. Key ingredients: model evaluation, diagnostics, selection, and building
   - Model evaluation and Model selection

Before we begin

- Feel free to visit my or Meike's Introducation to Biostatistics
- Meike's BSTA 511 page
- Nicky's BSTA 525 page

Learning Objectives

1. Identify important descriptive statistics and visualize data from a continuous variable
2. Identify important distributions that will be used in 512/612
3. Use our previous tools in 511 to estimate a parameter and construct a confidence interval
4. Use our previous tools in 511 to conduct a hypothesis test
5. Define error rates and power

Learning Objectives

1. Identify important descriptive statistics and visualize data from a continuous variable
2. Identify important distributions that will be used in 512/612
3. Use our previous tools in 511 to estimate a parameter and construct a confidence interval
4. Use our previous tools in 511 to conduct a hypothesis test
5. Define error rates and power

Some Basic Statistics "Talk"

- Random variable Y
  - Sample $Y_i, i = 1, \ldots, n$
- Summation:
  $\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + \ldots + Y_n$
- Product:
  $\prod_{i=1}^{n} Y_i = Y_1 \times Y_2 \times \ldots \times Y_n$

Descriptive Statistics: continuous variables

Measures of central tendency

- Sample mean

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median

Measures of variability (or dispersion)

- Sample variance
  - Average of the squared deviations from the sample mean
- Sample standard deviation

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

- IQR
  - Range from 1st to 3rd quartile

Descriptive Statistics: continuous variables (R code)

Measures of central tendency

- Sample mean

```r
mean( sample )
```

- Median

```r
median( sample )
```

Measures of variability (or dispersion)

- Sample variance

```r
var( sample )
```

- Sample standard deviation

```r
sd( sample )
```

- IQR

```r
IQR( sample )
```

- Or all together!!

```r
dds.discr %>% get_summary_stats(age)
```

```
# A tibble: 1 × 13
  variable      n   min   max median    q1    q3   iqr   mad  mean    sd    se
  <fct>     <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 age        1000     0    95     18    12    26    14  10.4  22.8  18.5 0.584
# ℹ 1 more variable: ci <dbl>
```

Data visualization

- Using the library `ggplot2` to visualize data
- We will load the package:

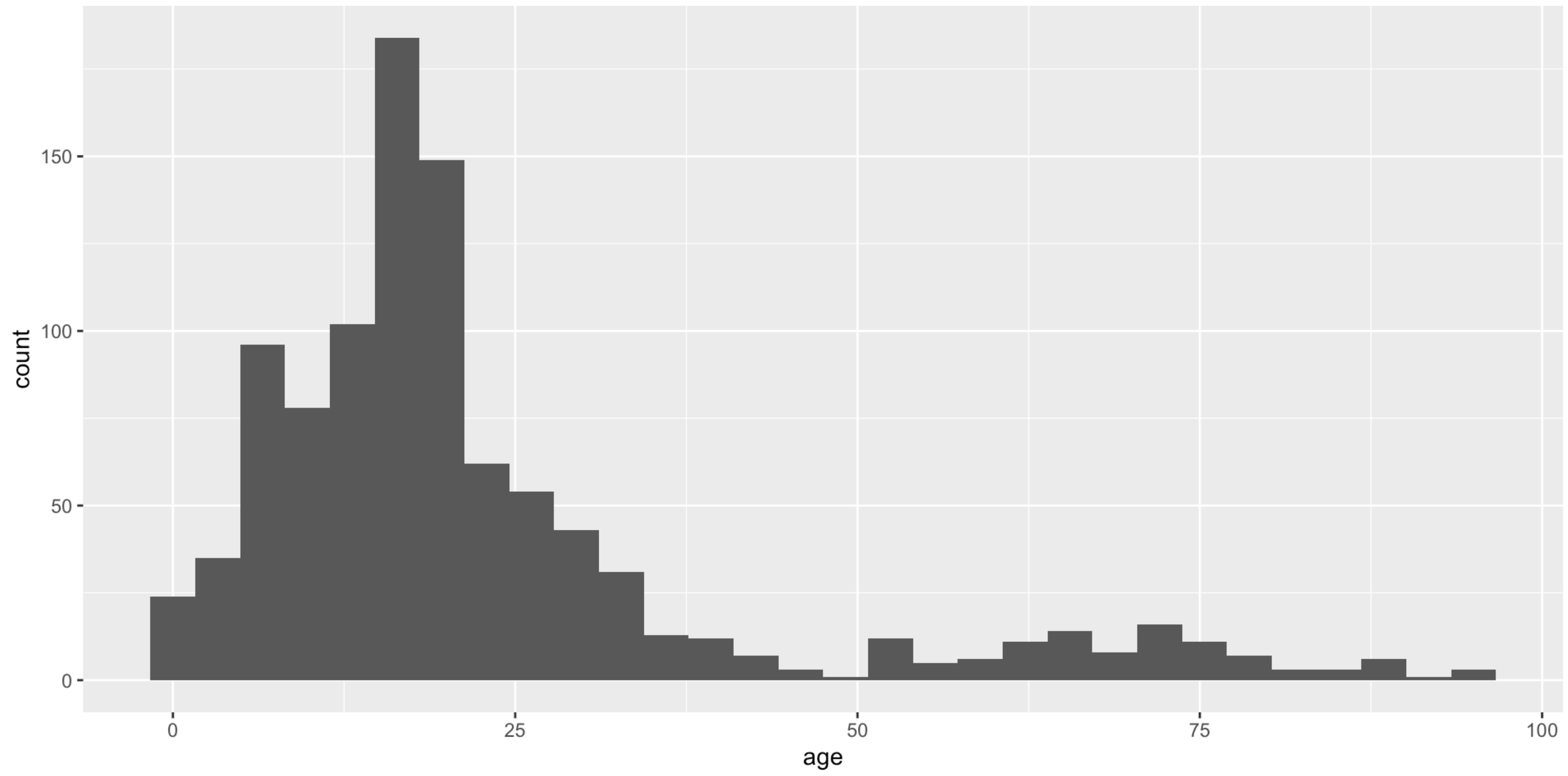
```
library(ggplot2)
```

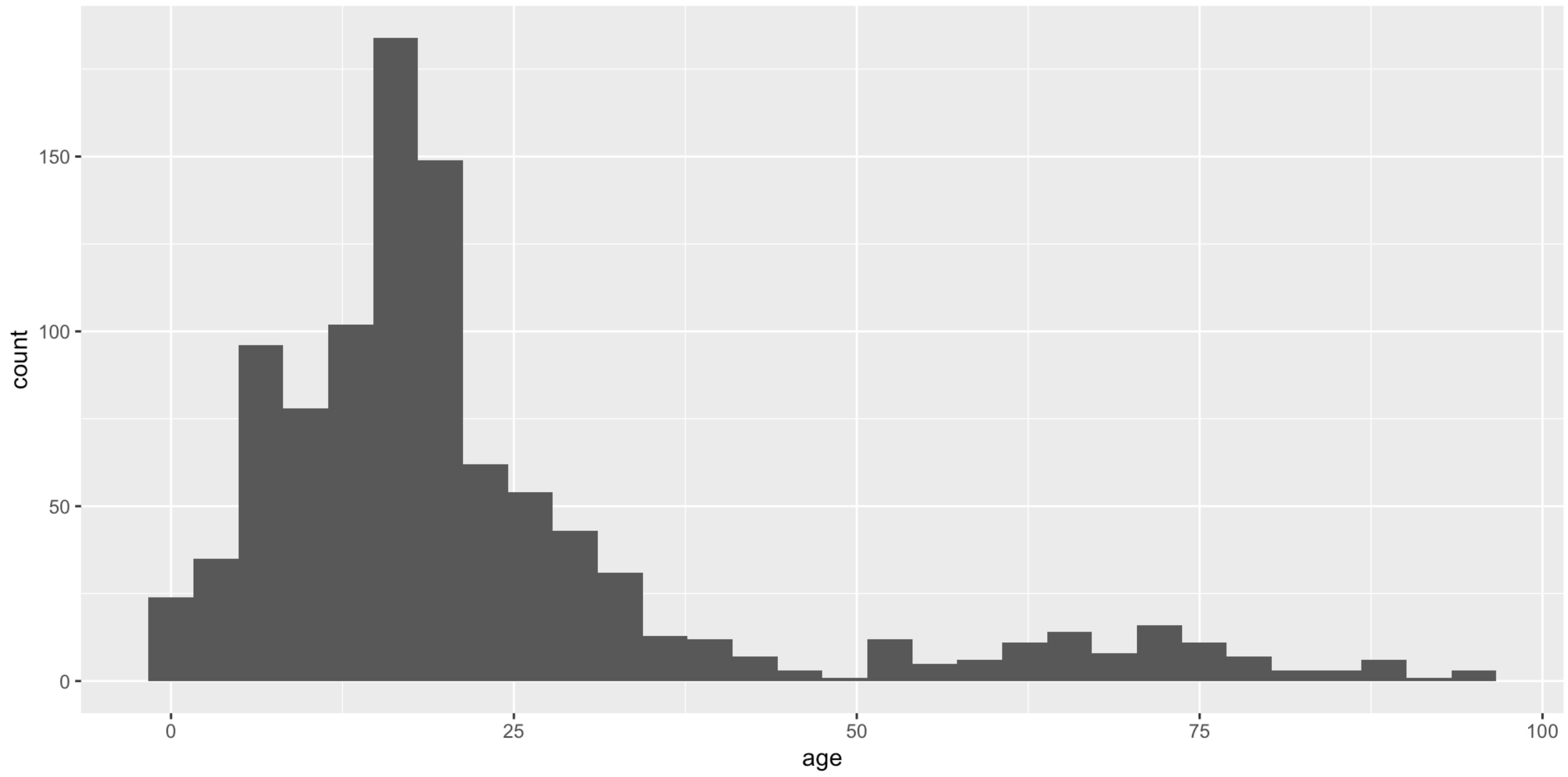
Histogram using ggplot2

We can make a basic graph for a continuous variable:

```r
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_histogram()
```

```r
ggplot() +
  geom_histogram(data = dds.discr,
                 aes(x = age))
```



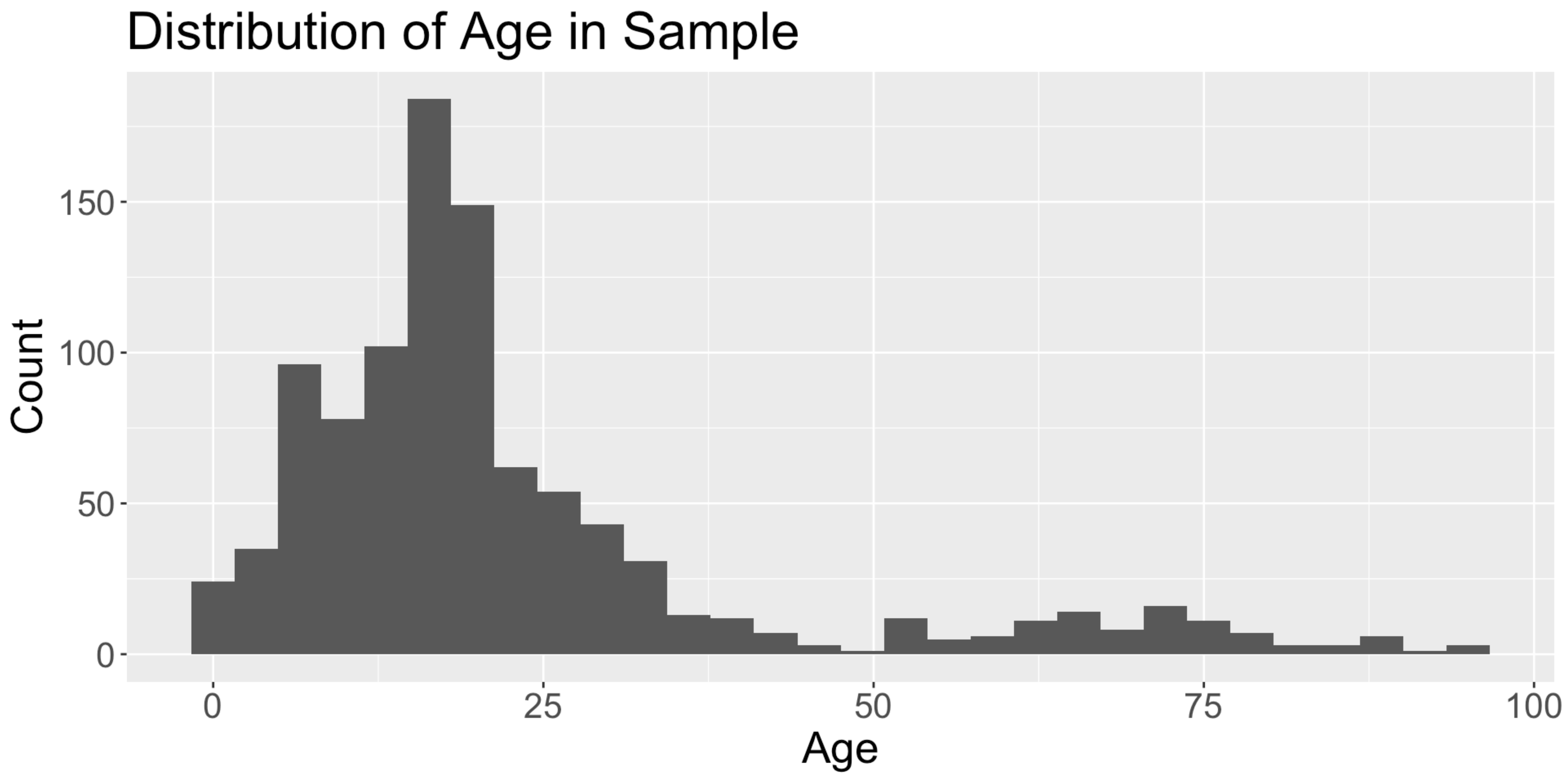


Some more information on histograms using ggplot2

Spruced up histogram using `ggplot2`

We can make a more formal, presentable graph:

```r
ggplot(data = dds.discr,
       aes(x = age)) +
  geom_histogram() +
  theme(text = element_text(size=20)) +
  labs(x = "Age",
       y = "Count",
       title = "Distribution of Age in Sample")
```

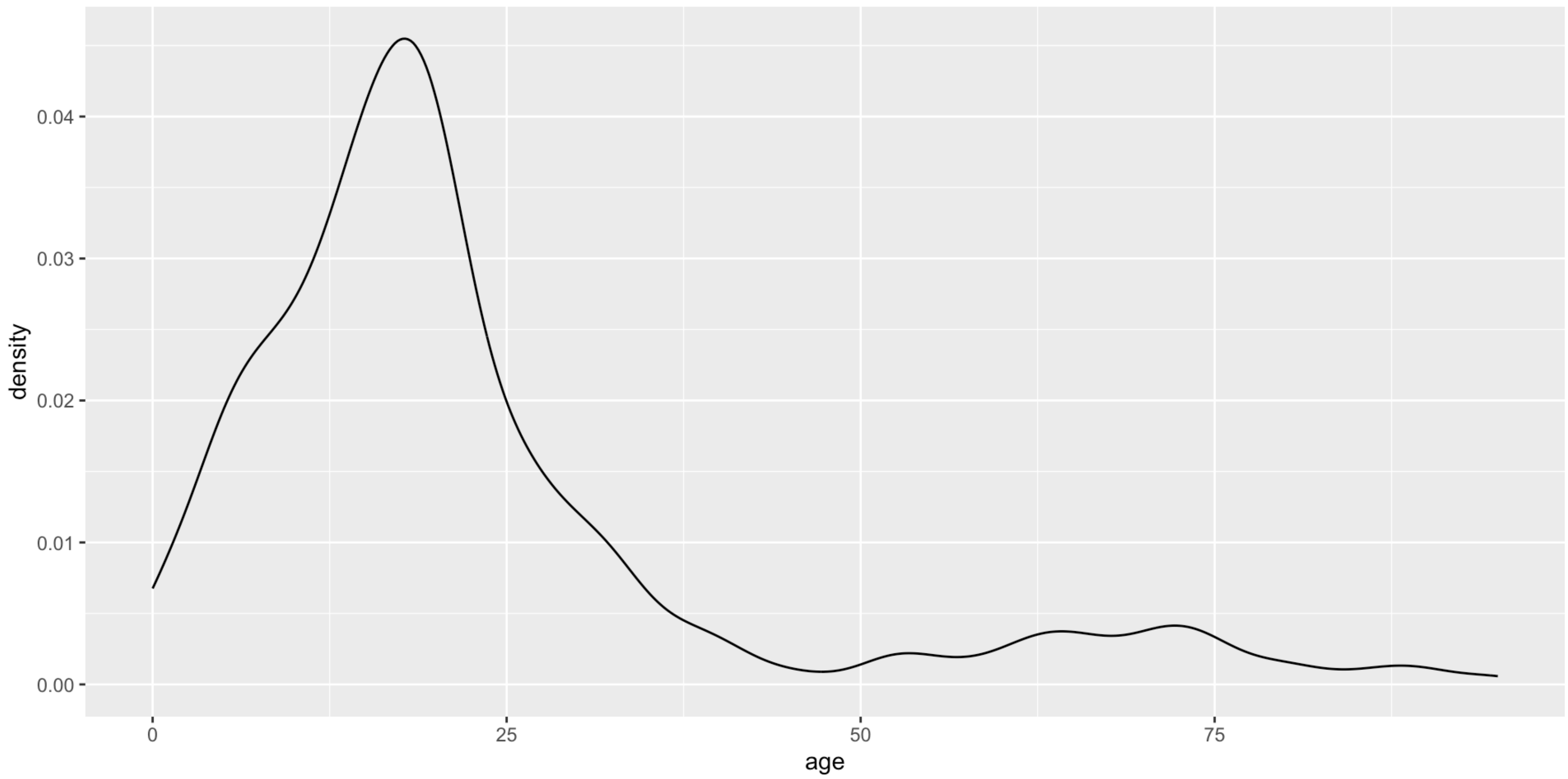


I would like you to turn in homework, labs, and project reports with graphs like these.

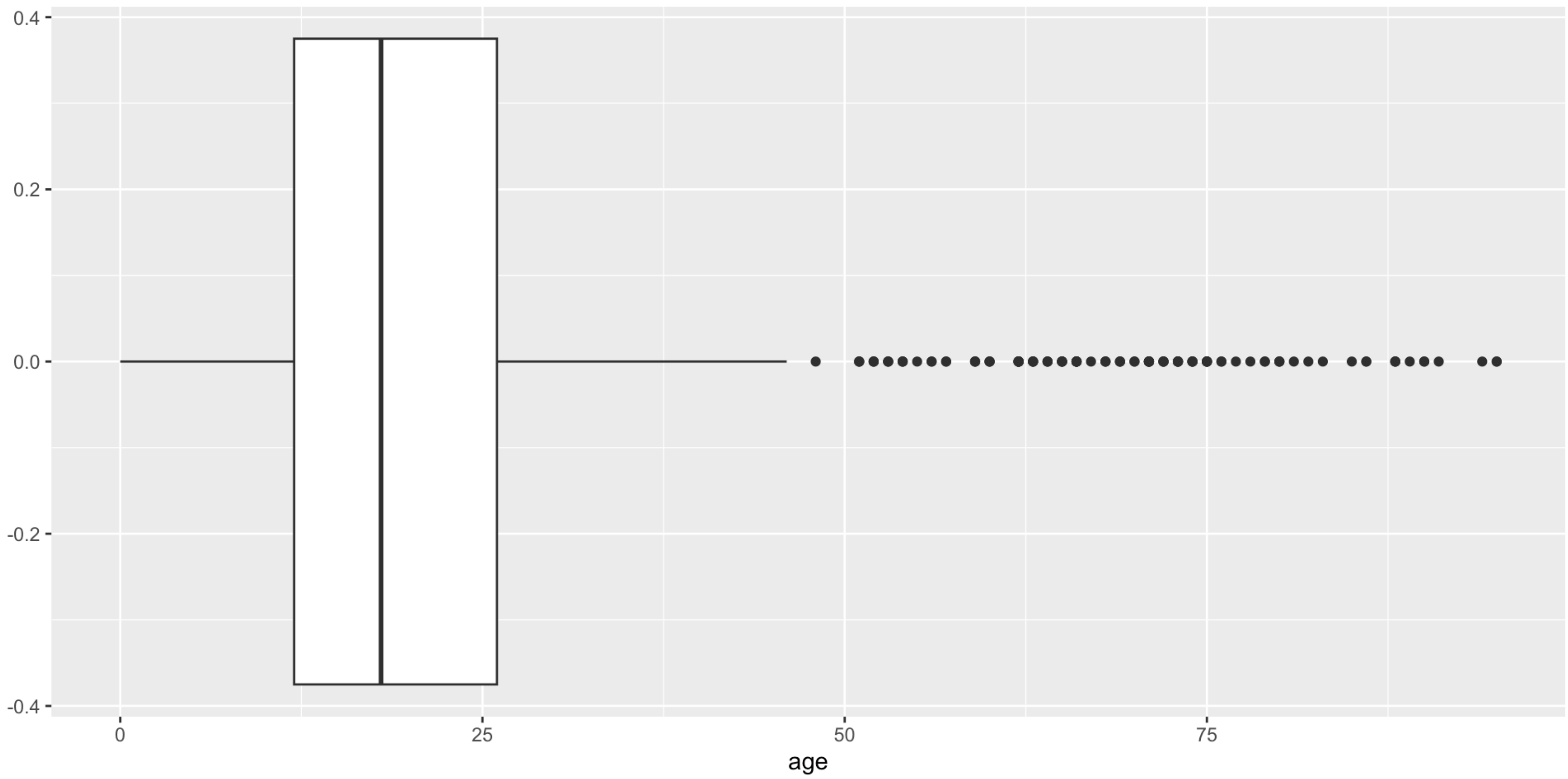
Other basic plots from `ggplot2`

We can also make a density and boxplot for the continuous variable with `ggplot2`

```
ggplot(data = dds.discr,
        aes(x = age)) +
  geom_density()
```




```
ggplot(data = dds.discr,
        aes(x = age)) +
  geom_boxplot()
```



Learning Objectives

1. Identify important descriptive statistics and visualize data from a continuous variable
2. Identify important distributions that will be used in 512/612
3. Use our previous tools in 511 to estimate a parameter and construct a confidence interval
4. Use our previous tools in 511 to conduct a hypothesis test
5. Define error rates and power

Distributions that will be used in this class

- Normal distribution
- Chi-square distribution
- Student’s t distribution
- F distribution

Normal Distribution

- Where did we see this?
  - Basically everywhere! Think Central Limit Theorem

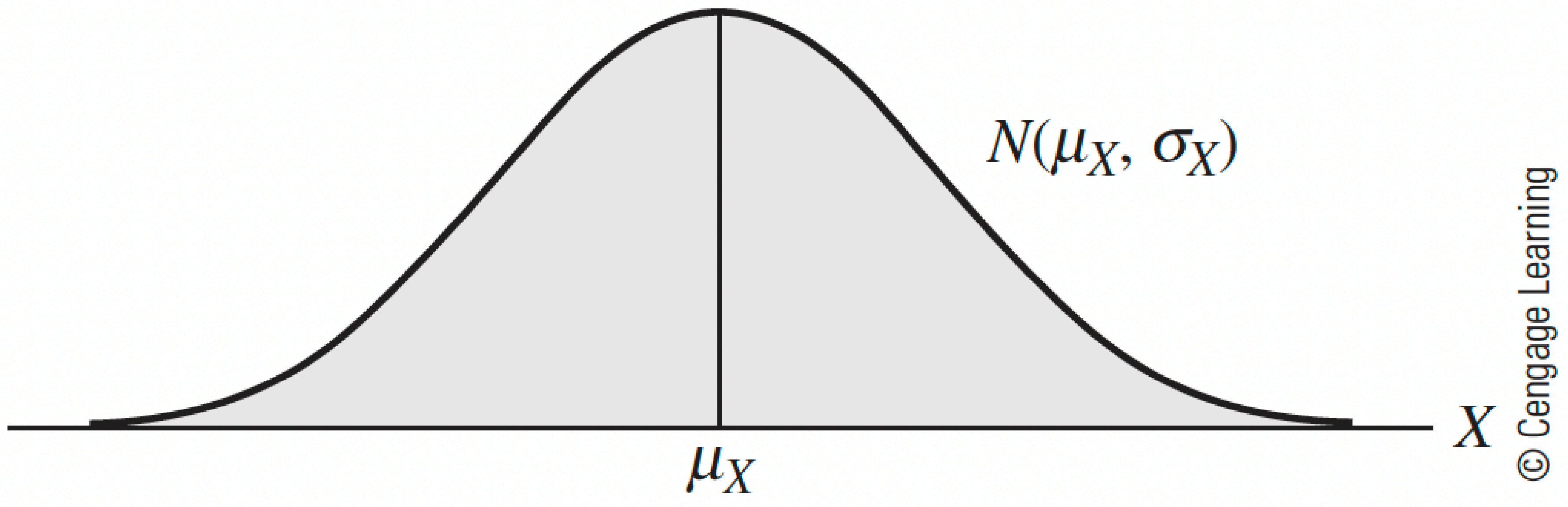


FIGURE 3.4 A normal distribution

- Notation: $Y \sim N(\mu, \sigma^2)$
- Arguably the most important distribution in statistics
- If we know $E(Y) = \mu, Var(Y) = \sigma^2$ then
  - 2/3 of Y's distribution lies within 1 σ of μ
  - 95% is within $\mu \pm 2\sigma$
  - $>$ 99% lies within $\mu \pm 3\sigma$

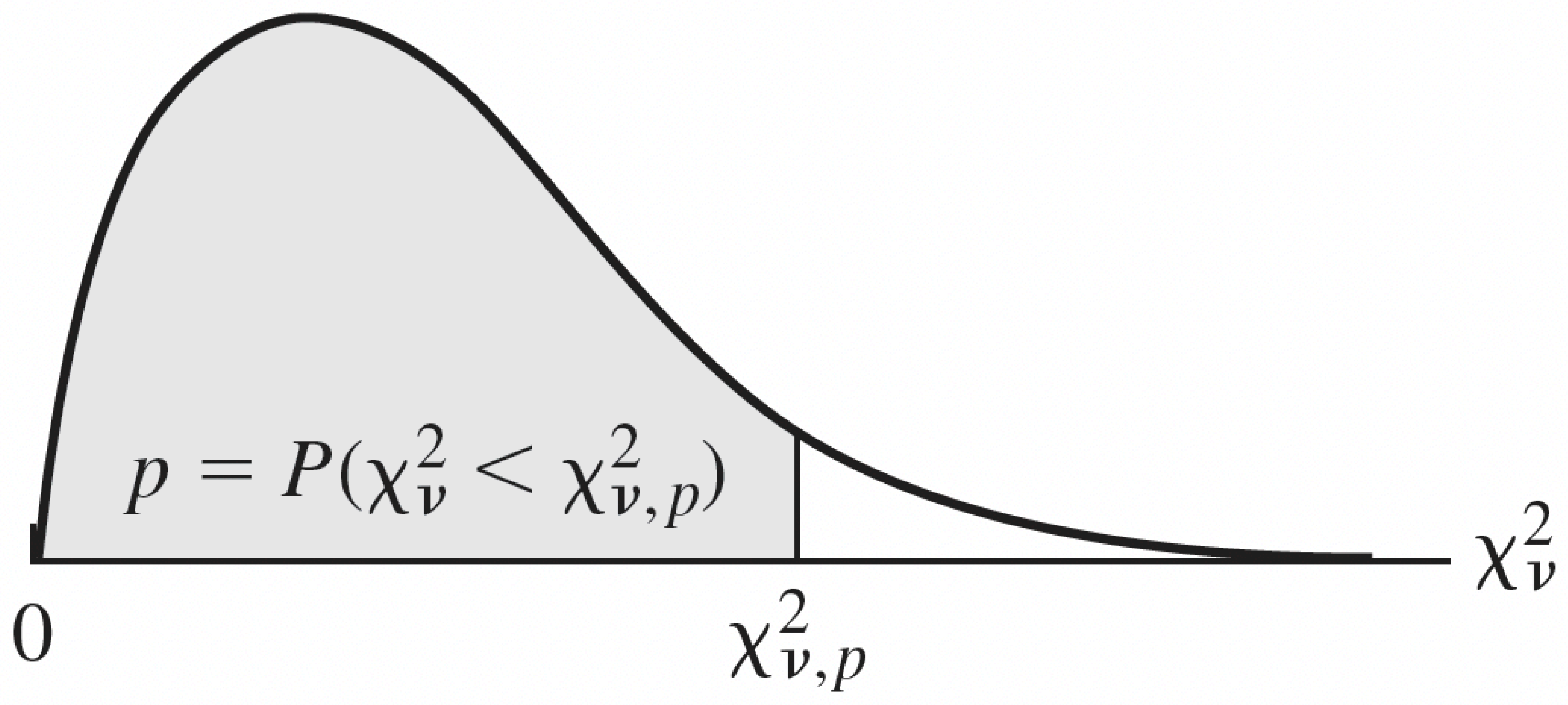
- Linear combinations of Normal's are Normal
  e.g., $(aY + b) \sim \mathrm{N}(a\mu + b,\ a^2\sigma^2)$
- Standard normal: $Z = \frac{Y-\mu}{\sigma} \sim \mathrm{N}(0, 1)$

Chi-squared distribution

- Where did we see this?
  - Hypothesis test if two categorical variables were independent

- Notation: $X \sim \chi^2_{df}$ OR $X \sim \chi^2_{\nu}$
  - Degrees of freedom (df): $df = n - 1$
  - X takes on only positive values

- If $Z_i \sim \mathrm{N}(0,1)$, then $Z_i^2 \sim \chi^2_1$
  - A standard normal distribution squared is the Chi squared distribution with df of 1.



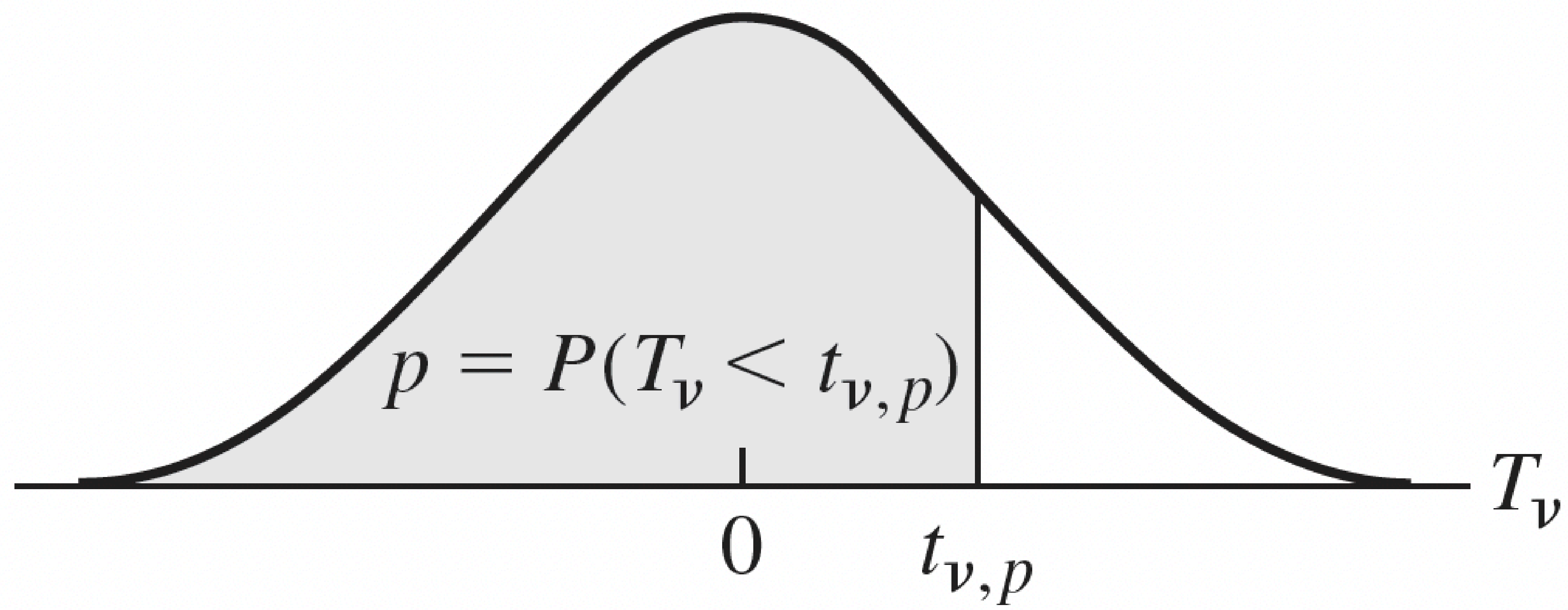
(b) χ^2 distribution

Student's t Distribution

- Where did we see this?
  - Inference of means: single sample, paired, two independent samples

- Notation: $T \sim t_{df}$ OR $T \sim t_{n-1}$
  - Degrees of freedom (df): $df = n - 1$
  - $T = \dfrac{\overline{x} - \mu_x}{\dfrac{s}{\sqrt{n}}} \sim t_{n-1}$

- In linear modeling, used for inference on individual regression parameters
  - Think: our estimated coefficients ($\hat{\beta}$)



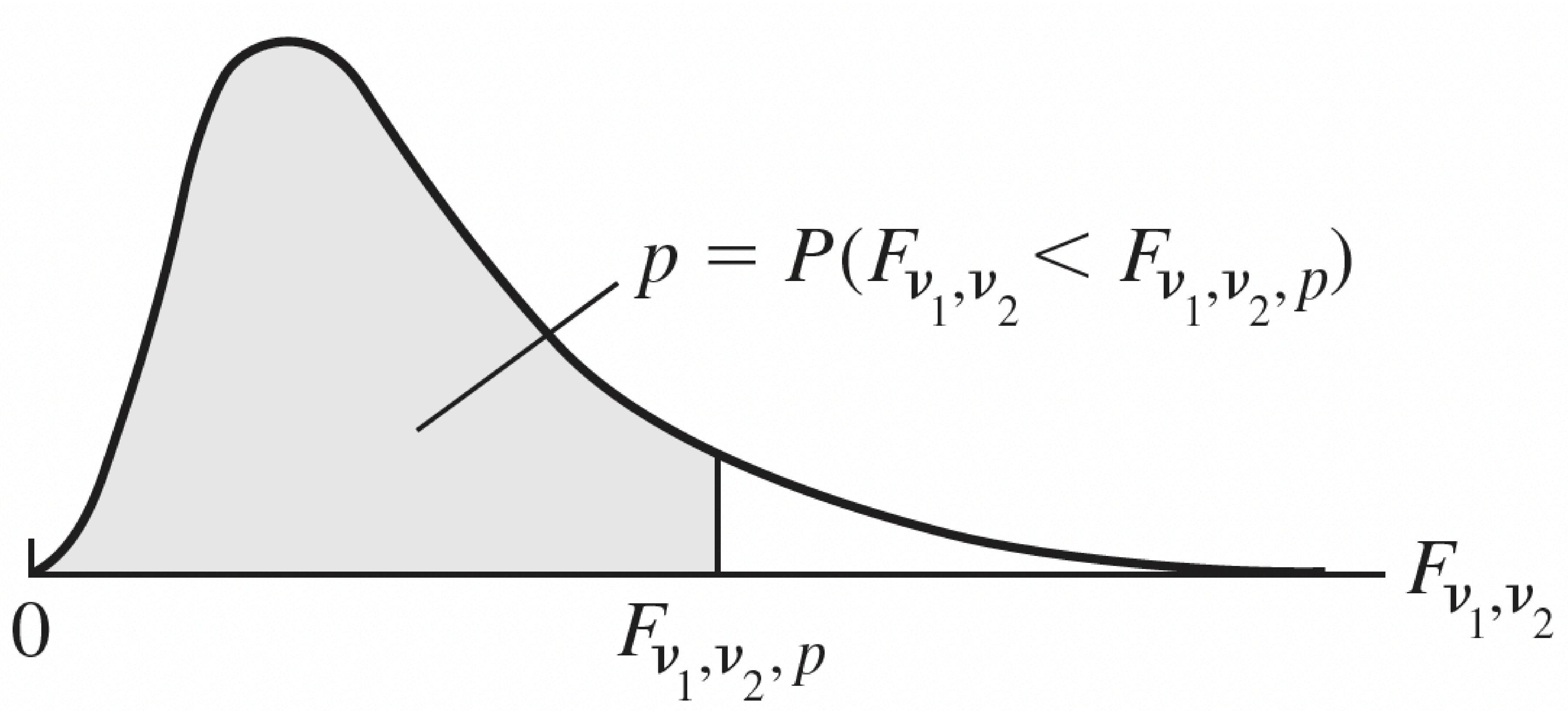
(a) Student's *t* distribution

F-Distribution

- Where did we see this?
  - Inference for 2+ means: ANOVA test
- Model ratio of sample variances (and is a ratio of Chi-squared RVs)
- If $X_1^2 \sim \chi^2_{df1}$ and $X_2^2 \sim \chi^2_{df2}$, where $X_1^2 \perp X_2^2$, then:

$$\frac{X_1^2/df1}{X_2^2/df2} \sim F_{df1,df2}$$

- Important relationship with t distribution: $T^2 \sim F_{1,\nu}$
  - The square of a t-distribution with $df = \nu$
  - is an F-distribution with numerator df ($df_1 = 1$) and denominator df ($df_2 = \nu$)



(c) *F* distribution

R code for probability distributions

Here is a site with the various probability distributions and their R code.

- It also includes practice with R code to see what each function outputs

Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial (including Bernoulli)	pbinom	qbinom	dbinom	rbinom
Birthday	pbirthday	qbirthday		
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Discrete Uniform	sample			
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Multinomial			dmultinom	rmultinom
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Kolmogorov-Smirnov Test Statistic	psmirnov	qsmirnov		rsmirnov
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Continuous Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank
Wishart				rWishart

Learning Objectives

1. Identify important descriptive statistics and visualize data from a continuous variable
2. Identify important distributions that will be used in 512/612
3. Use our previous tools in 511 to estimate a parameter and construct a confidence interval
4. Use our previous tools in 511 to conduct a hypothesis test
5. Define error rates and power

Confidence interval for one mean

The confidence interval for population mean μ:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- where t^* is the critical value for the 95% (or other percent) corresponding to the t-distribution and dependent on $df = n - 1$

We can use R to find the critical t-value, t^*

For example the critical value for the 95% CI with $n = 10$ subjects is...

```
qt(0.975, df=9)
```

```
[1] 2.262157
```

- Recall, that as the df increases, the t-distribution converges towards the Normal distribution

Confidence interval for one mean

The confidence interval for population mean μ:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

- where t^* is the critical value for the 95% (or other percent) corresponding to the t-distribution and dependent on $df = n - 1$

We can use R to find the critical t-value, t^*

For example the critical value for the 95% CI with $n = 10$ subjects is…

```
qt(0.975, df=9)
```

```
[1] 2.262157
```

- Recall, that as the df increases, the t-distribution converges towards the Normal distribution

We can also use `t.test` in R to calculate the confidence interval if we have a dataset.

```
t.test(dds.discr$age)
```

```

	One Sample t-test

data:  dds.discr$age
t = 39.053, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal
to 0
95 percent confidence interval:
 21.65434 23.94566
sample estimates:
mean of x
     22.8
```

Confidence interval for two independent means

The confidence interval for difference in independent population means, μ_1 and μ_2:

$$\overline{x}_1 - \overline{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- where t^* is the critical value for the 95% (or other percent) corresponding to the t-distribution and dependent on $df = n_1 + n_2 - 2$

- Please check out my notes on this if you'd like: https://nwakim.github.io/F24_EPI_525/schedule.html
  - It's under Lesson 13

Here's a decent source for other R code for tests in 511

Website from UCLA

Learning Objectives

1. Identify important descriptive statistics and visualize data from a continuous variable
2. Identify important distributions that will be used in 512/612
3. Use our previous tools in 511 to estimate a parameter and construct a confidence interval
4. Use our previous tools in 511 to conduct a hypothesis test
5. Define error rates and power

Reference: Steps in a Hypothesis Test

1. Check the **assumptions**
    - What sampling distribution are you using? What assumptions are required for it?
2. Set the **level of significance** α
3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**
    - In symbols and/or in words
    - Alternative: one- or two-sided?
4. Calculate the **test statistic**.
5. Calculate the **p-value** based on the observed test statistic and its sampling distribution
6. Write a **conclusion** to the hypothesis test
    - Do we reject or fail to reject H_0?
    - Write a conclusion in the context of the problem

Another view: Steps in a Hypothesis Test

1. Check the assumptions regarding the properties of the underlying variable(s) being measured that are needed to justify use of the testing procedure under consideration.
2. State the null hypothesis H_0 and the alternative hypothesis H_A.
3. Specify the significance level α.
4. Specify the test statistic to be used and its distribution under H_0.

Critical region method

5. Form the decision rule for rejecting or not rejecting H_0 (i.e., specify the rejection and nonrejection regions for the test, based on both H_A and α).
6. Compute the value of the test statistic from the observed data.

***p*-value method**

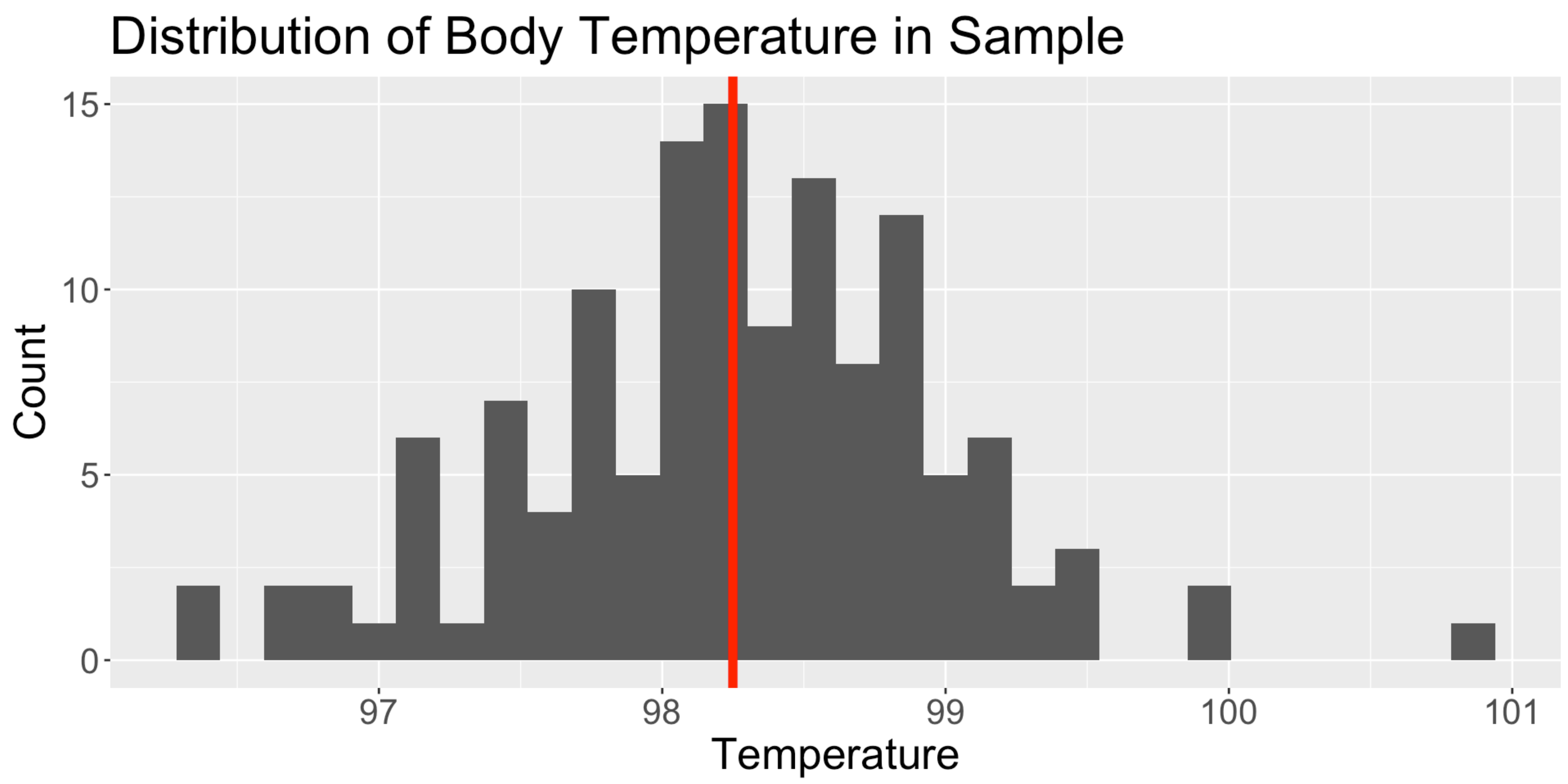
5. Compute the value of the test statistic from the observed data.
6. Calculate the p-value

7. Draw conclusions regarding rejection or nonrejection of H0.

Example: one sample t-test

```r
BodyTemps = read.csv("data/BodyTemperatures.csv")

ggplot(data = BodyTemps,
       aes(x = Temperature)) +
  geom_histogram() +
  theme(text = element_text(size=20)) +
  labs(x = "Temperature", y = "Count",
       title = "Distribution of Body Temperature in Sample") +
  geom_vline(aes(xintercept = mean(BodyTemps$Temperature, na.rm = T)),
             color = "red", linewidth = 2)
```



Reference: what does it all look like together?

Example of hypothesis test based on the 1992 JAMA data

Is there evidence to support that the population mean body temperature is different from 98.6°F?

1. **Assumptions:** The individual observations are independent and the number of individuals in our sample is 130. Thus, we can use CLT to approximate the sampling distribution.

2. Set $\alpha = 0.05$

3. **Hypothesis:**

$$\begin{aligned} H_0 &: \mu = 98.6 \\ \text{vs. } H_A &: \mu \neq 98.6 \end{aligned}$$

4-5. Test statistic and p-value

▶ Code

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
98.24923	-5.454823	2.410632e-07	129	98.122	98.37646	One Sample t-test	two.sided

6. **Conclusion:** We reject the null hypothesis. The average body temperature in the sample was 98.25°F (95% CI 98.12, 98.38°F), which is discernibly different from 98.6°F (p-value < 0.001).

How did we get the 95% CI?

- The `t.test` function can help us answer this, and give us the needed information for both approaches.

```r
BodyTemps = read.csv("data/BodyTemperatures.csv")

t.test(x = BodyTemps$Temperature,
        # alternative = "two-sided",
        mu = 98.6)
```

```
	One Sample t-test

data:  BodyTemps$Temperature
t = -5.4548, df = 129, p-value = 2.411e-07
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
 98.12200 98.37646
sample estimates:
mean of x 
 98.24923 
```

Learning Objectives

1. Identify important descriptive statistics and visualize data from a continuous variable
2. Identify important distributions that will be used in 512/612
3. Use our previous tools in 511 to estimate a parameter and construct a confidence interval
4. Use our previous tools in 511 to conduct a hypothesis test
5. Define error rates and power

Outcomes of our hypothesis test

TABLE 3.1 Outcomes of hypothesis testing

Hypothesis Chosen	True State of Nature: H_0	True State of Nature: H_A
H_0	Correct decision	False negative decision (Type II error)
H_A	False positive decision (Type I error)	Correct decision

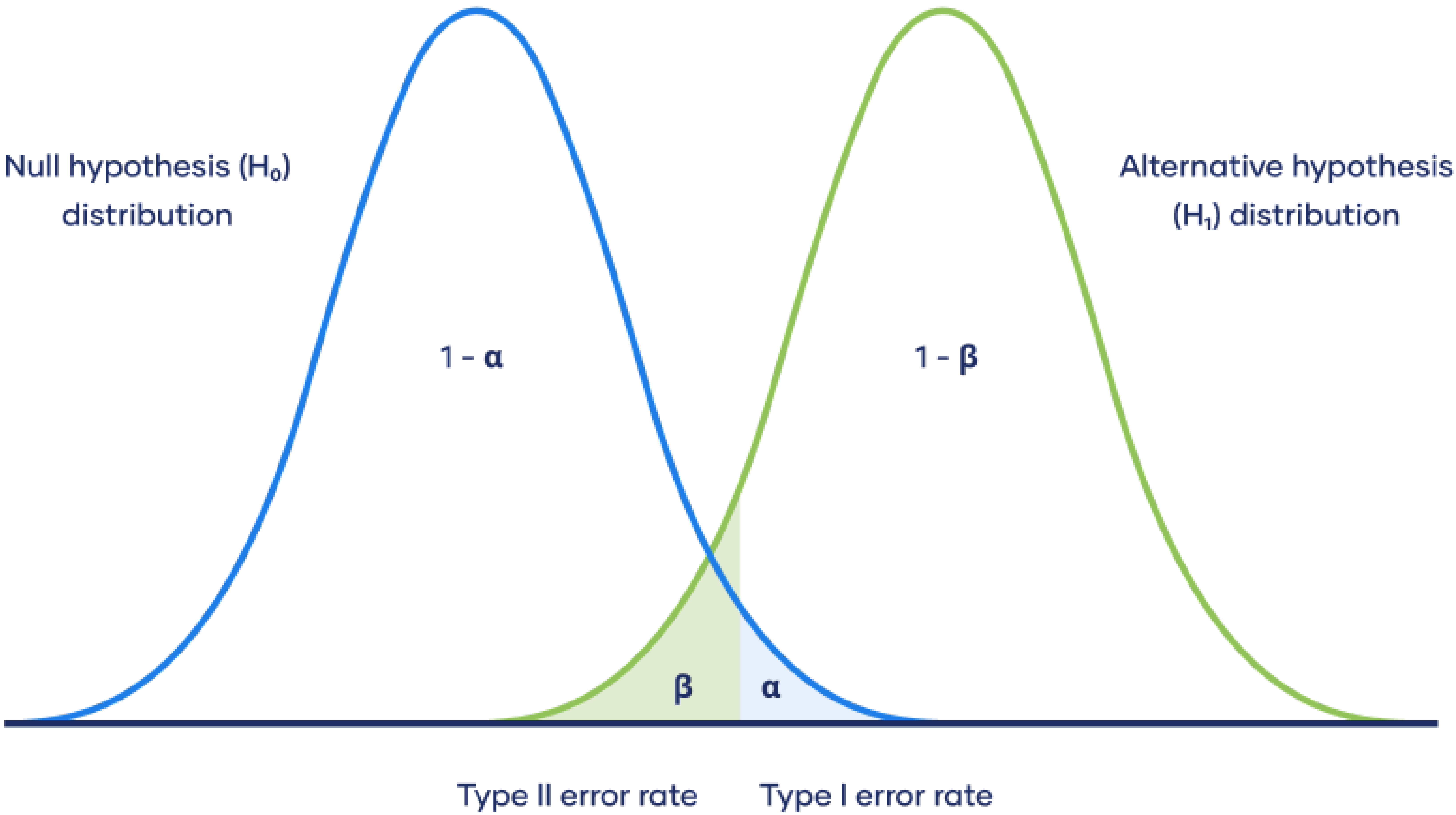
Prabilities of outcomes

- Type 1 error is α
  - The probability that we falsely reject the null hypothesis (but the null is true!!)
- Power is $1 - \beta$
  - The probability of correctly rejecting the null hypothesis

TABLE 3.2 Probabilities of outcomes of hypothesis testing

Hypothesis Chosen	True State of Nature	
	H_0	H_A
H_0	$1 - \alpha$	β
H_A	α	$1 - \beta$

What I think is the most intuitive way to look at it



Do your exit ticket!!

- Don’t forget to go online and fill it out!
  - This will count as your attendance

- I look forward to the quarter with you!