Review

EPSY 630 - Statistics II

Jason Bryer, Ph.D.

April 20, 2021

Announcements and Agenda

Sign-up for a presentation time for either May 4th or 11th on this Google sheet.

Next week we will discuss Bayesian Analysis. This will be the last topic.

Today's Agenda:

1 Review the statistical procedures we learned this semester

2 Questions

3 One minute papers

Tutoring Data

An observational study that examined the effects of tutoring services on students' grades in English courses in an online college.

```
data(tutoring, package = 'TriMatch')
tutoring <- tutoring %>%
    mutate(treat2 = treat %in%
               c('Treat1', 'Treat2'),
           Pass = Grade >= 2)
```

```
str(tutoring)
```

```
## 'data.frame':    1142 obs. of  19 variables:
##  $ treat     : Factor w/ 3 levels "Control","Treat1",..: 1 1
##  $ Course    : chr  "ENG*201" "ENG*201" "ENG*201" "ENG*201"
##  $ Grade     : int  4 4 4 4 4 3 4 3 0 4 ...
##  $ Gender    : Factor w/ 2 levels "FEMALE","MALE": 1 1 1 1 1
##  $ Ethnicity : Factor w/ 3 levels "Black","Other",..: 2 3 3
##  $ Military  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ ESL       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ EdMother  : int  3 5 1 3 2 3 4 4 3 6 ...
##  $ EdFather  : int  6 6 1 5 2 3 4 4 2 6 ...
##  $ Age       : num  48 49 53 52 47 53 54 54 59 40 ...
##  $ Employment: int  3 3 1 3 1 3 3 3 1 3 ...
##  $ Income    : num  9 9 5 5 5 9 6 6 1 8 ...
##  $ Transfer  : num  24 25 39 48 23 ...
##  $ GPA       : num  3 2.72 2.71 4 3.5 3.55 3.57 3.57 3.43 2.
##  $ GradeCode : chr  "A" "A" "A" "A" ...
##  $ Level     : Factor w/ 2 levels "Lower","Upper": 1 1 1 1 1
##  $ ID        : int  377 882 292 215 252 265 1016 282 39 911
##  $ treat2    : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ Pass      : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

Types of Data

Quantitative variables represent amounts of things (e.g. the number of trees in a forest). Types of quantitative variables include:

- **Continuous** (a.k.a ratio variables): represent measures and can usually be divided into units smaller than one (e.g. 0.75 grams).
- **Discrete** (a.k.a integer variables): represent counts and usually can't be divided into units smaller than one (e.g. 1 tree).

Categorical variables represent groupings of things (e.g. the different tree species in a forest). Types of categorical variables include:

- **Ordinal**: represent data with an order (e.g. rankings).
- **Nominal**: represent group names (e.g. brands or species names).
- **Binary**: represent data with a yes/no or 1/0 outcome (e.g. win or lose).

Descriptive Statistics

Quantitative Variables

Measures of center:

- Mean (`mean`)
- Median (`median`)
- Mode

Measures of spread:

- Variance (`var`)
- Standard deviation (`sd`)
- Interquartile range (`IQR`)

Plots

- Histogram
- Density
- Box plot

Qualitative Variables

- Contingency table (`table`)
- Proportional table (`prop.table`)

Plots

- Bar plot
- Mosaic plot

Central Limit Theorem

The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

where SE represents the **standard error**, which is defined as the standard deviation of the sampling distribution. In most cases σ is not known, so use s.

Central Limit Theorem (cont.)

Consider the following population...

```
N <- 1000000
pop <- rbeta(N,2,20)
ggplot(data.frame(x = pop), aes(x = x)) + geom_density()
```

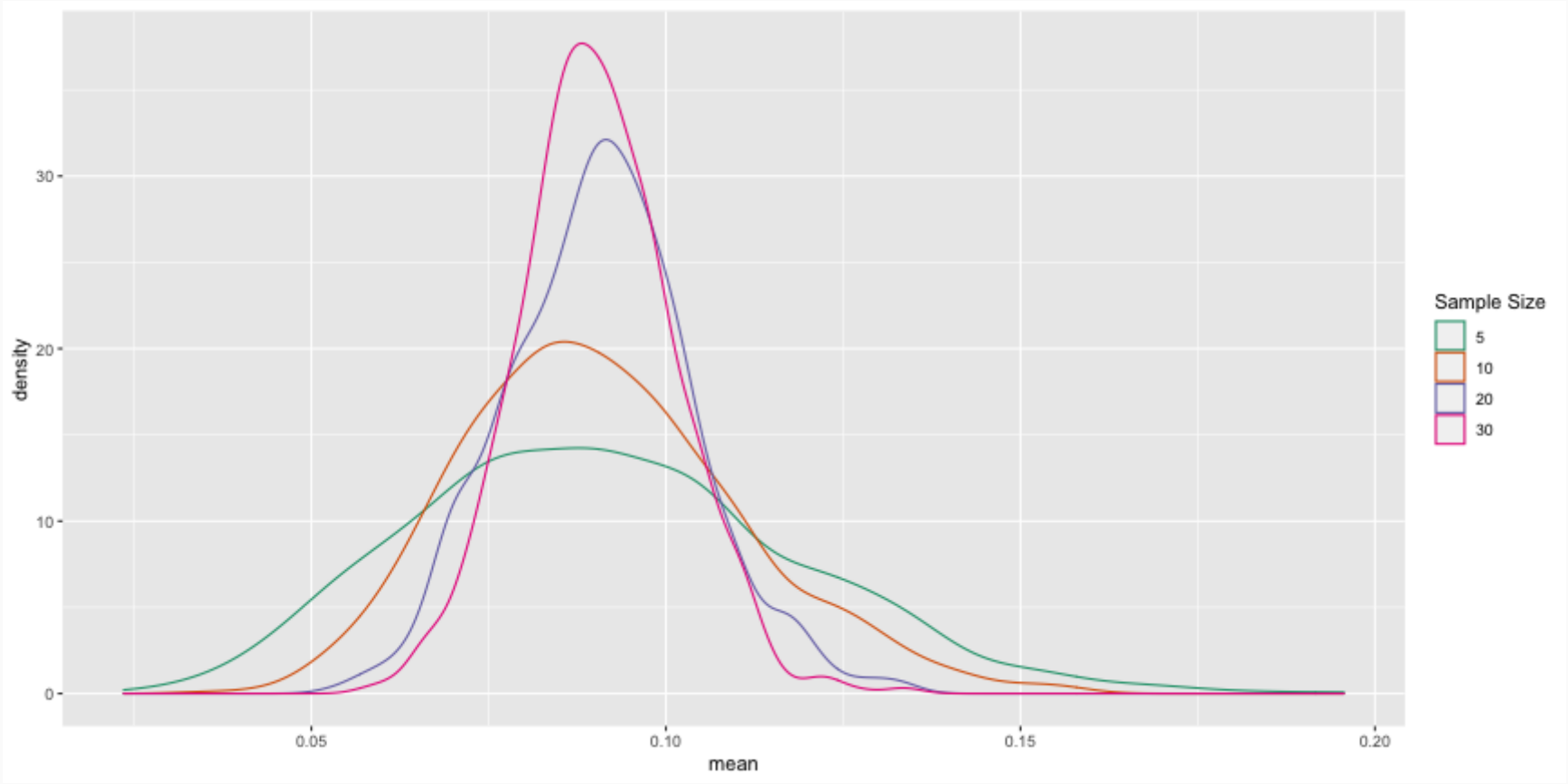
Estimating Sampling Distributions

Here, we will estimate 4 sampling distributions by taking 1,000 random samples from the population with sample sizes of 5, 10, 20, and 30 each.

```
n_samples <- 1000
df <- tibble(n = rep(c(5, 10, 20, 30), each = n_samples),
             mean = NA_real_)
for(i in 1:nrow(df)) {
    df[i,]$mean <- mean(sample(pop, size = df[i,]$n))
}
```

Sampling Distributions

```
ggplot(df) + geom_density(aes(x = mean, color = factor(n))) +
    scale_color_brewer('Sample Size', type = 'qual', palette = 2)
```



Null Hypothesis Testing

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Regression

> Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').
>
> Wikipedia

Regression problems take on the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

For *i* predictor variables where β_0 is the intercept and β_i is the slope for predictor i.

Statistical Assumptions

Statistical tests make some common assumptions about the data they are testing:

1. **Independence of observations** (aka no autocorrelation): The observations/variables you include in your test are not related (for example, multiple measurements of a single test subject are not independent, while measurements of multiple different test subjects are independent).

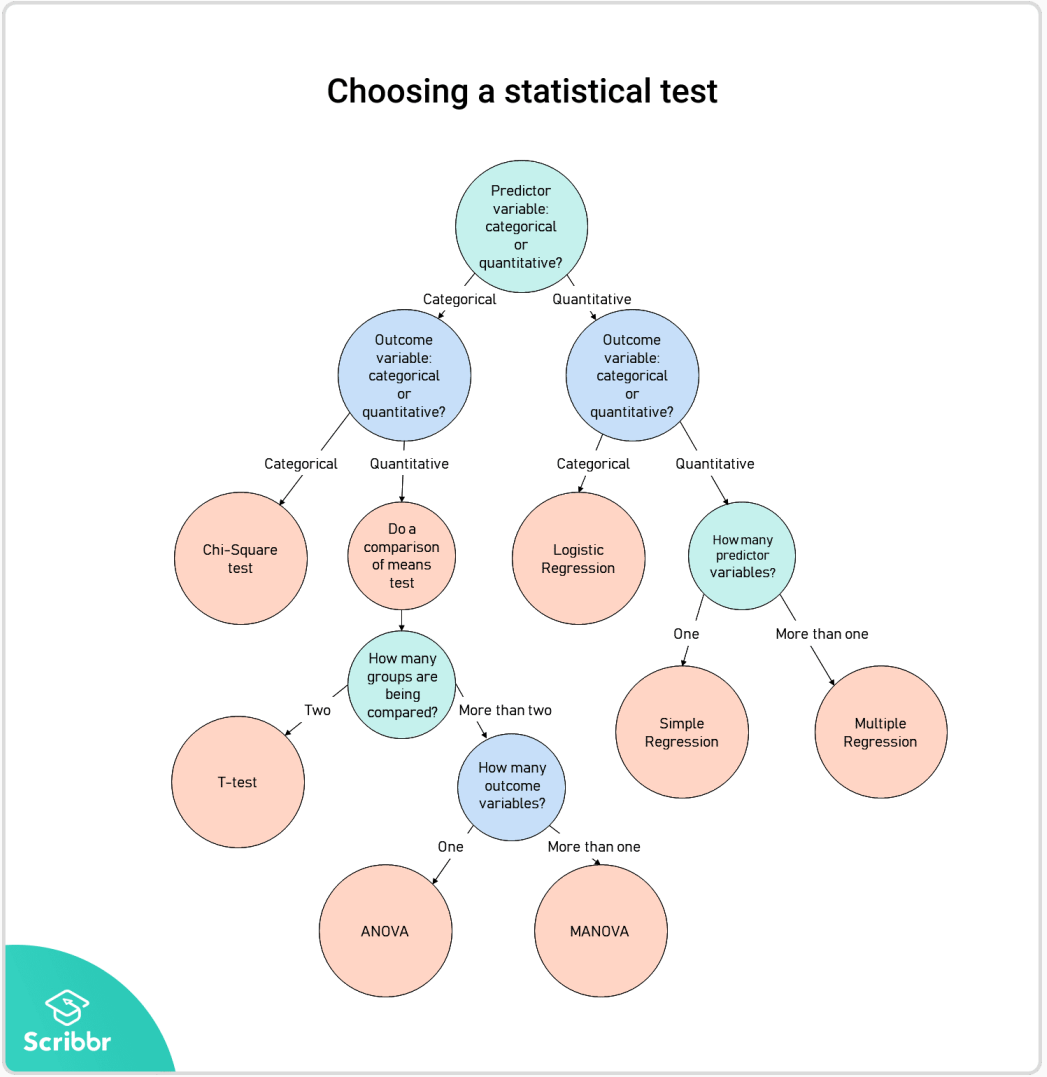
2. **Homogeneity of variance**: the variance within each group being compared is similar among all groups. If one group has much more variation than others, it will limit the test's effectiveness.

3. **Normality of data**: the data follows a normal distribution (aka a bell curve). This assumption applies only to quantitative data.

Choosing a statistical test

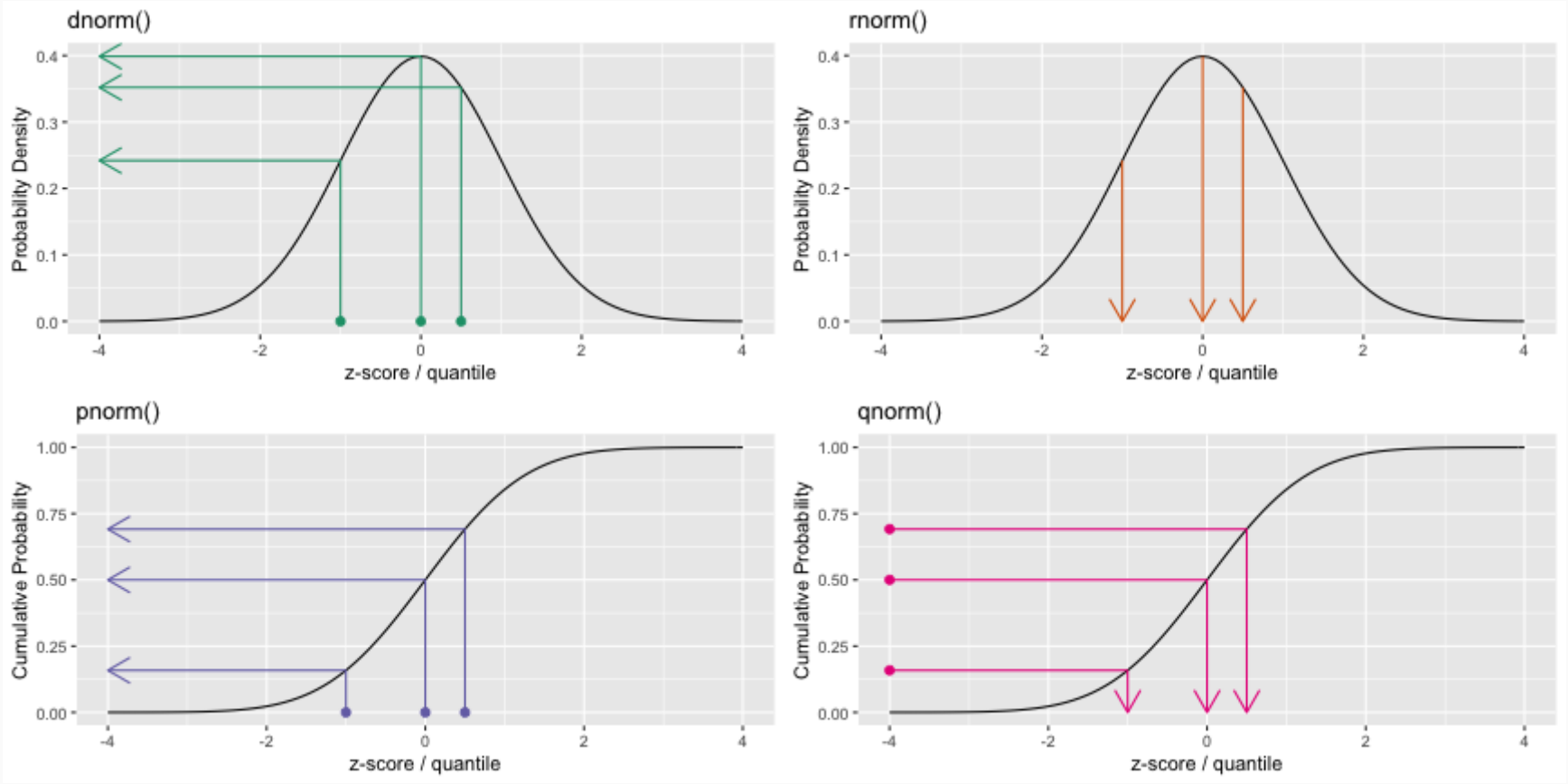
Test	Predictor Variable	Outcome variable	R function
Paired t-test	Categorical, 1	Quantitative	`t.test`
Independent t-test	Categorical, 1	Quantitative	`t.test`
Chi-Squared test	Categorical, 1 or more	Categorical	`chisq.test`
ANOVA	Categorical, 1 or more	Quantitative	`aov`
MANOVA	Categorical, 1 or more	Quantitative, 2 or more	`manova`
Correlation	Quantitative	Quantitative	`cor.test`
Linear regression	Quantitative	Quantitative	`lm`
Multiple regression	Any, 2 or more	Quantitative	`lm`
Logistic regression	Any, 1 or more	Categorical (dichotomous)	`glm(family=binomial(link='logit'))`

Choosing a statistical test



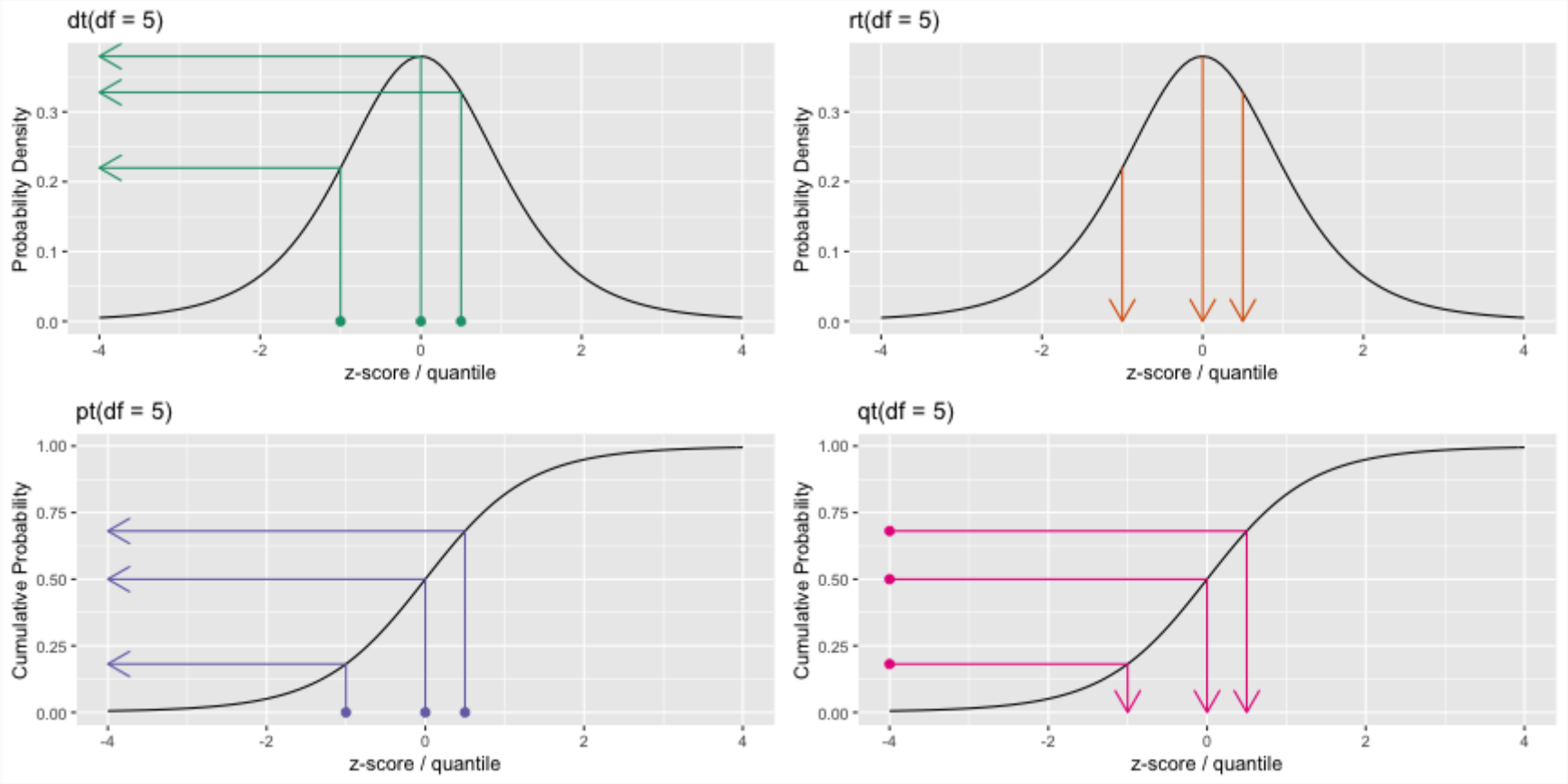
Normal Distribution

```
plot_distributions(dist = 'norm', xvals = c(-1, 0, 0.5), xmin = -4, xmax = 4)
```



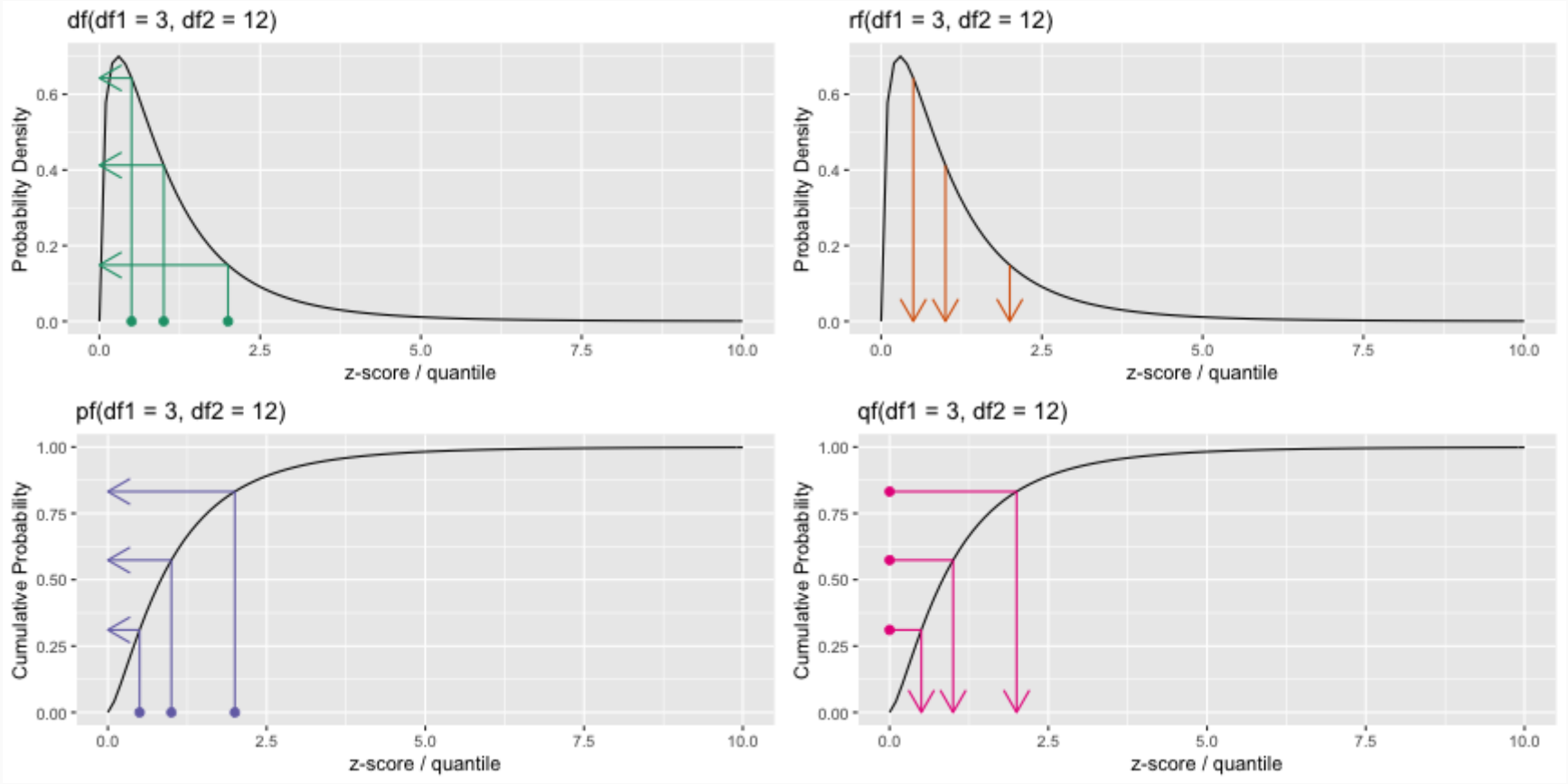
t Distribution

```
plot_distributions(dist = 't', xvals = c(-1, 0, 0.5), xmin = -4, xmax = 4,
                   args = list(df = 5))
```



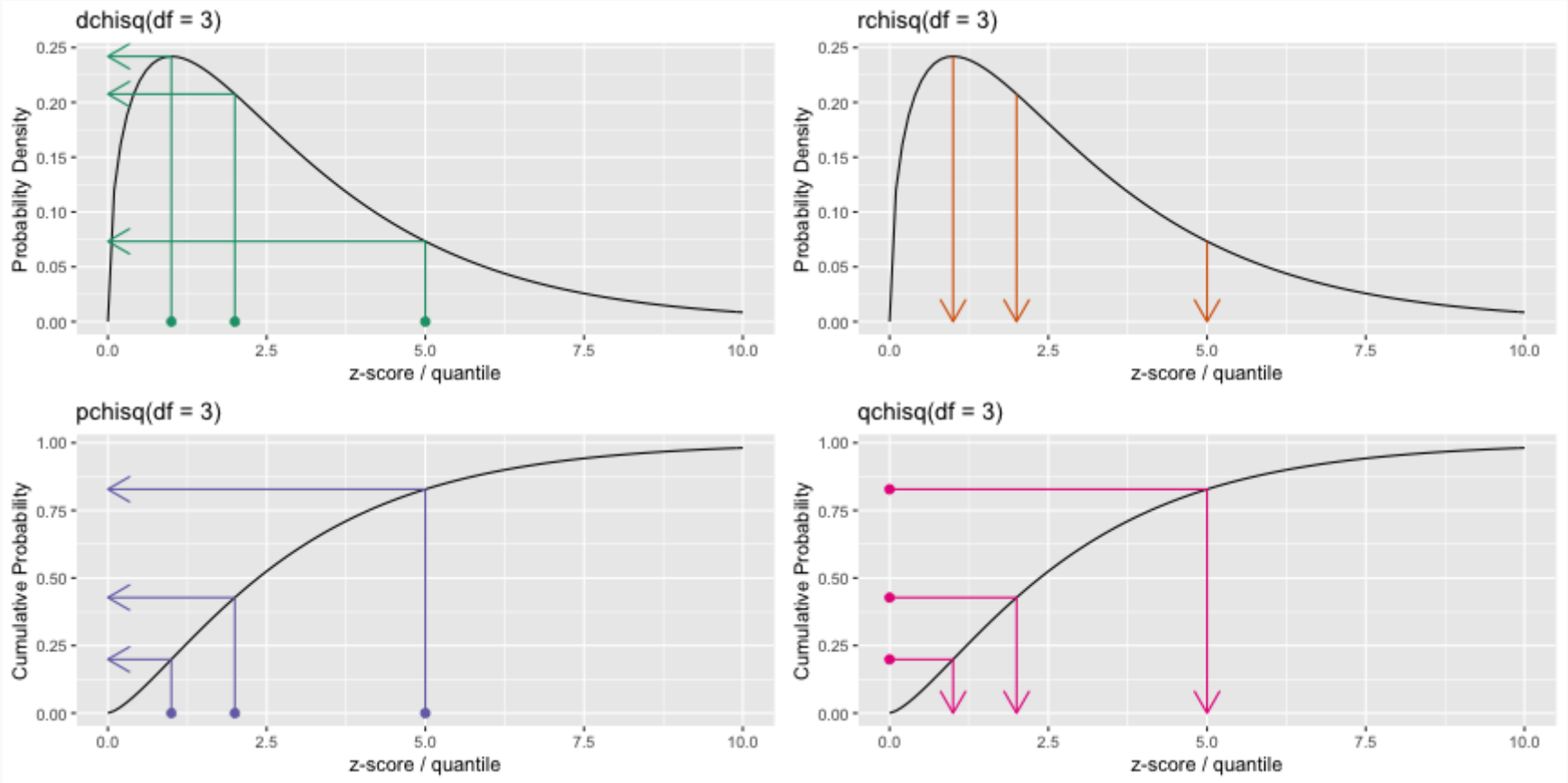
F Distribution

```
plot_distributions(dist = 'f', xvals = c(0.5, 1, 2), xmin = 0, xmax = 10,
                   args = list(df1 = 3, df2 = 12))
```



χ^2 Distribution

```
plot_distributions(dist = 'chisq', xvals = c(1, 2, 5), xmin = 0, xmax = 10,
                   args = list(df = 3))
```



Examples

Dependent (paired) *t*-test

RQ: Is there a difference educational attainment between mothers and fathers for students?

H_0: There is no difference in the educational attainment between mothers and fathers.

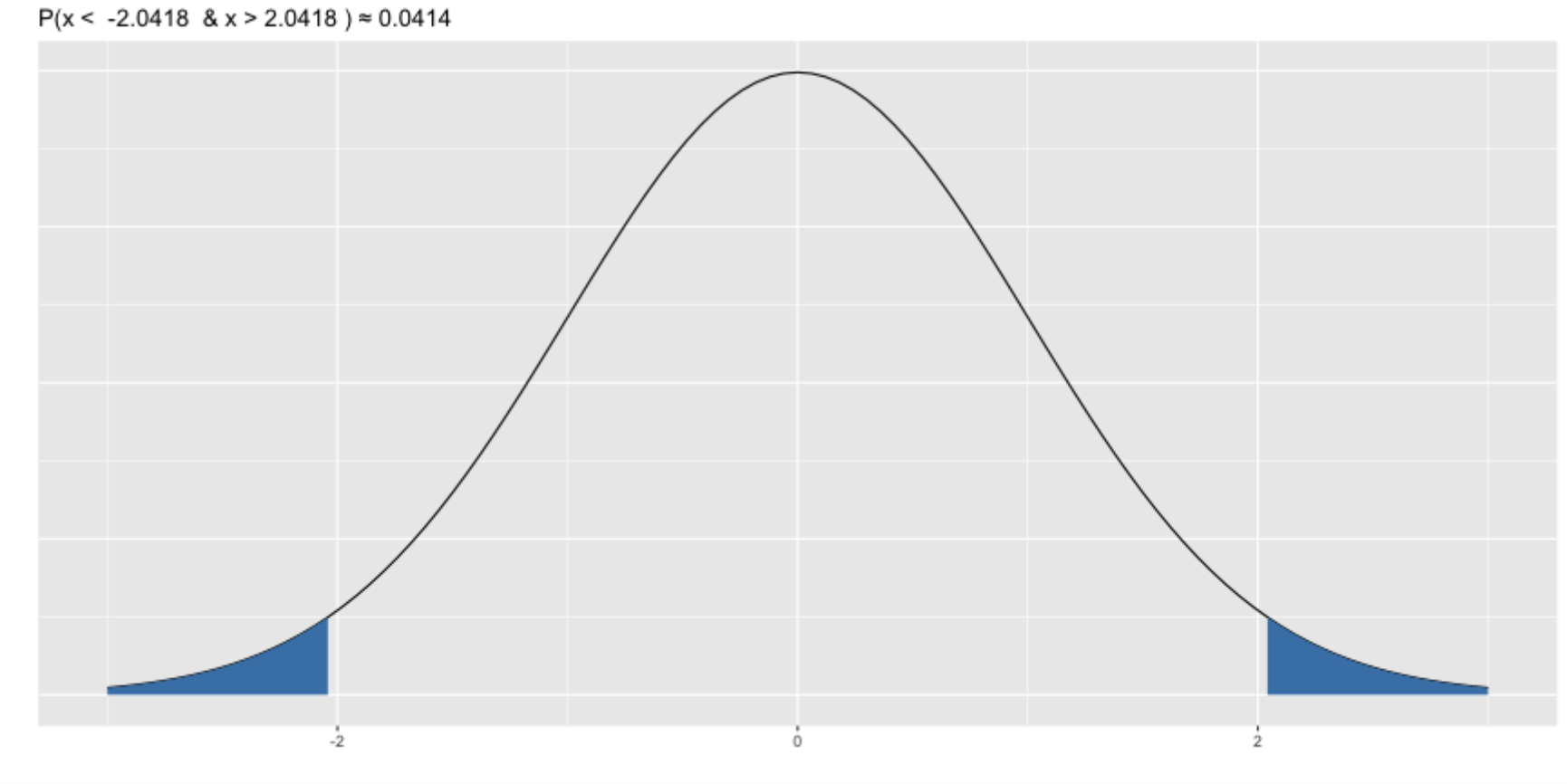
H_A: There is a difference in the educational attaintment between mothers and fathers.

```
t.test(tutoring$EdMother, tutoring$EdFather, paired = TRUE)
```

```
##
##     Paired t-test
##
## data:  tutoring$EdMother and tutoring$EdFather
## t = 2.0418, df = 1141, p-value = 0.0414
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.003934477 0.197466574
## sample estimates:
## mean of the differences
##               0.1007005
```

Dependent (paired) *t*-test

```
distribution_plot(dt, df = 1141, cv = 2.0418, limits = c(-3, 3), tails = 'two.sided')
```



Independent *t*-test

RQ: Is there a difference in GAP between military and civilian students?

H_0: There is no differences in GPA between military and civilian students?

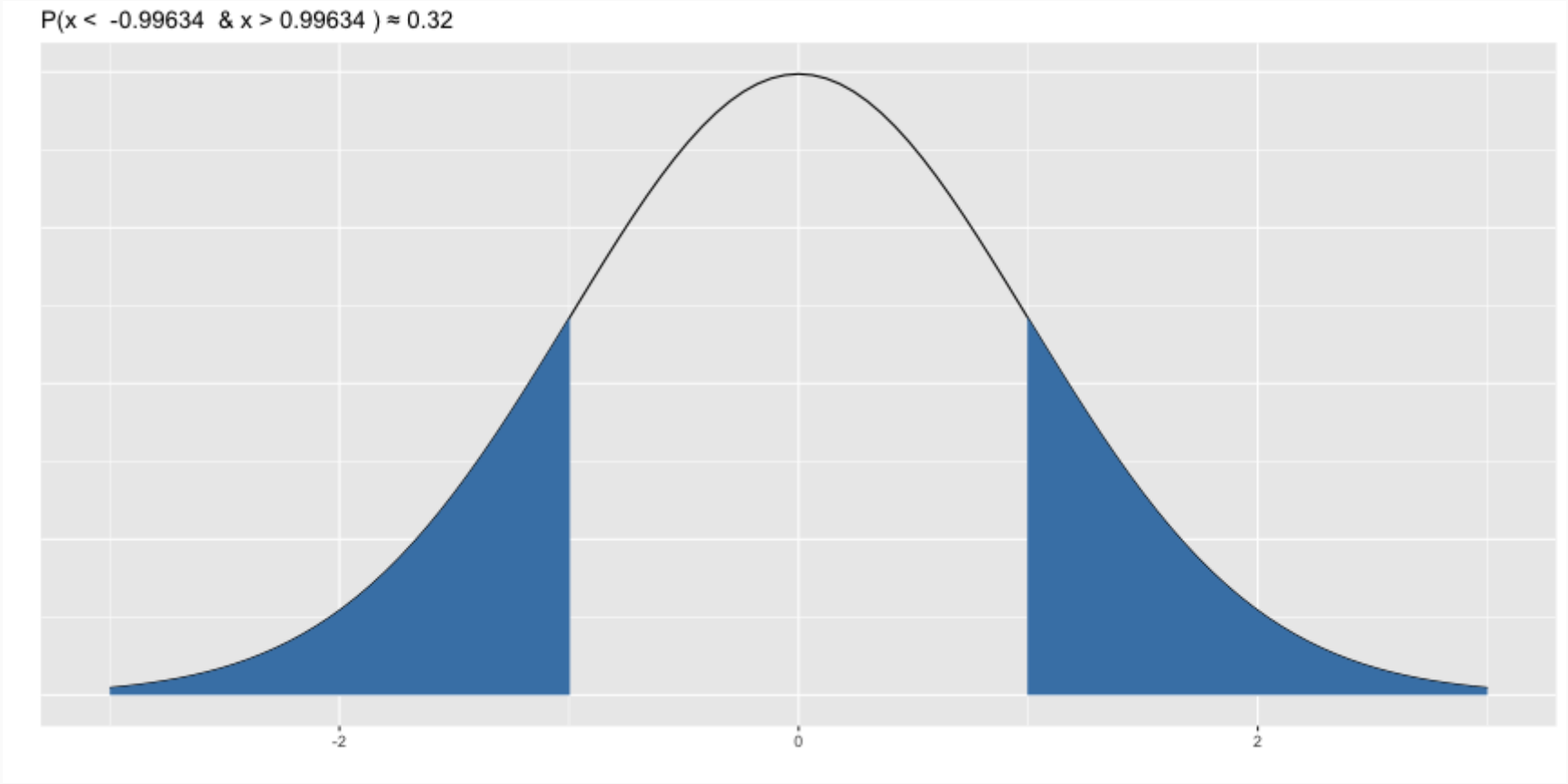
H_A: There is a difference in GPA between military and civilian students?

```
t.test(GPA ~ Military, data = tutoring)
```

```
##
##     Welch Two Sample t-test
##
## data:  GPA by Military
## t = 0.99634, df = 480.03, p-value = 0.3196
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04205162  0.12856489
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            3.179719            3.136462
```

Dependent (paired) *t*-test

```
distribution_plot(dt, df = 480, cv = 0.99634, limits = c(-3, 3), tails = 'two.sided')
```



χ^2 test

RQ: Is there a difference in the passing rate between students who used tutoring services and those who did not?

H_0: There is no difference in the passing rate by treatment.

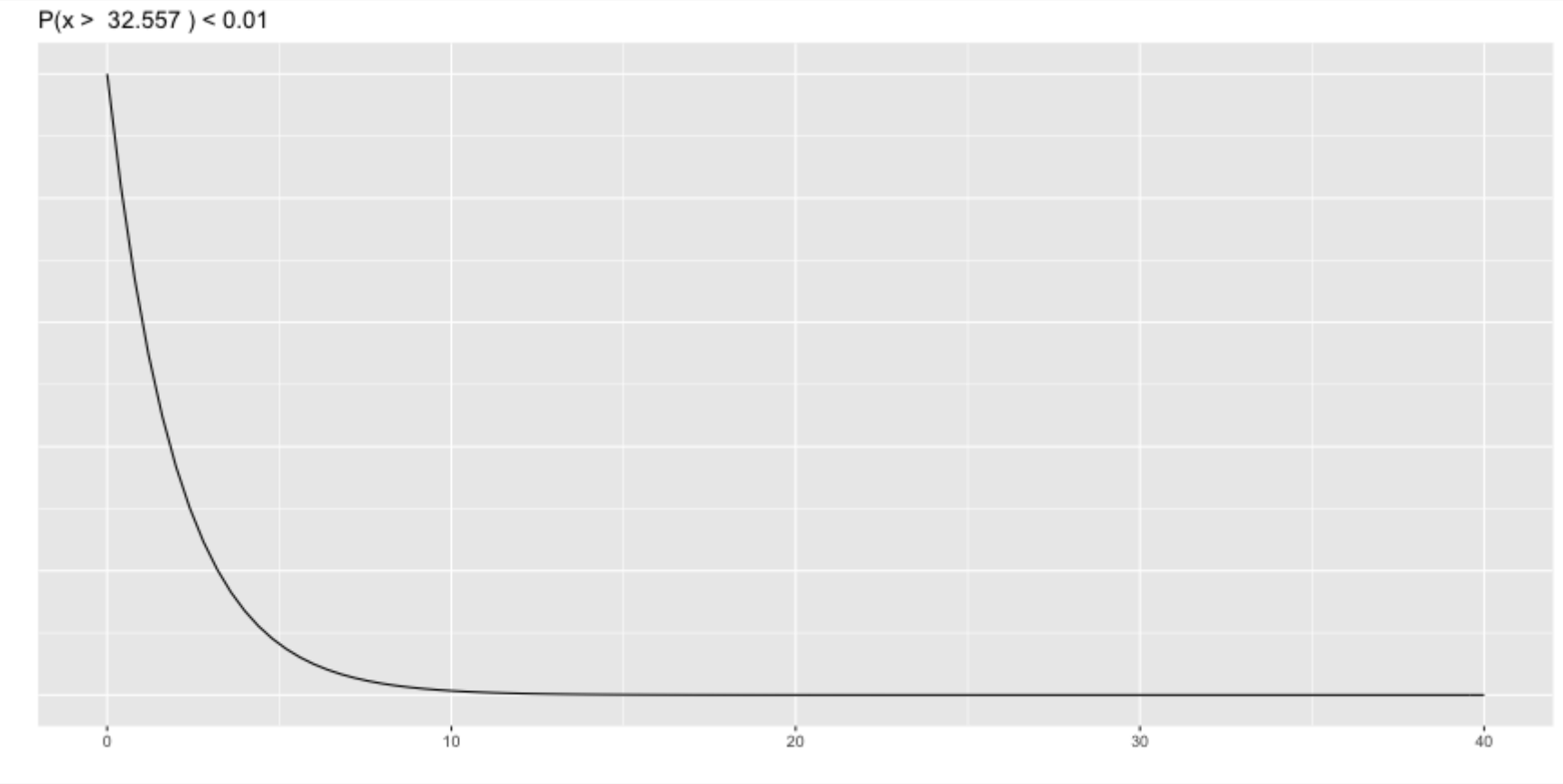
H_A: There is a difference in the passing rate by treatment.

```
chisq.test(tutoring$treat, tutoring$Pass)
```

```
##
##     Pearson's Chi-squared test
##
## data:  tutoring$treat and tutoring$Pass
## X-squared = 32.557, df = 2, p-value = 8.516e-08
```

χ^2 test

```
distribution_plot(dchisq, df = 2, cv = 32.557,
                  limits = c(0, 40), tails = 'greater')
```



Analysis of Variance (ANOVA)

RQ: Is there a difference in GPA by ethnicity?

H_0: The mean GPA is the same for all ethnicities.

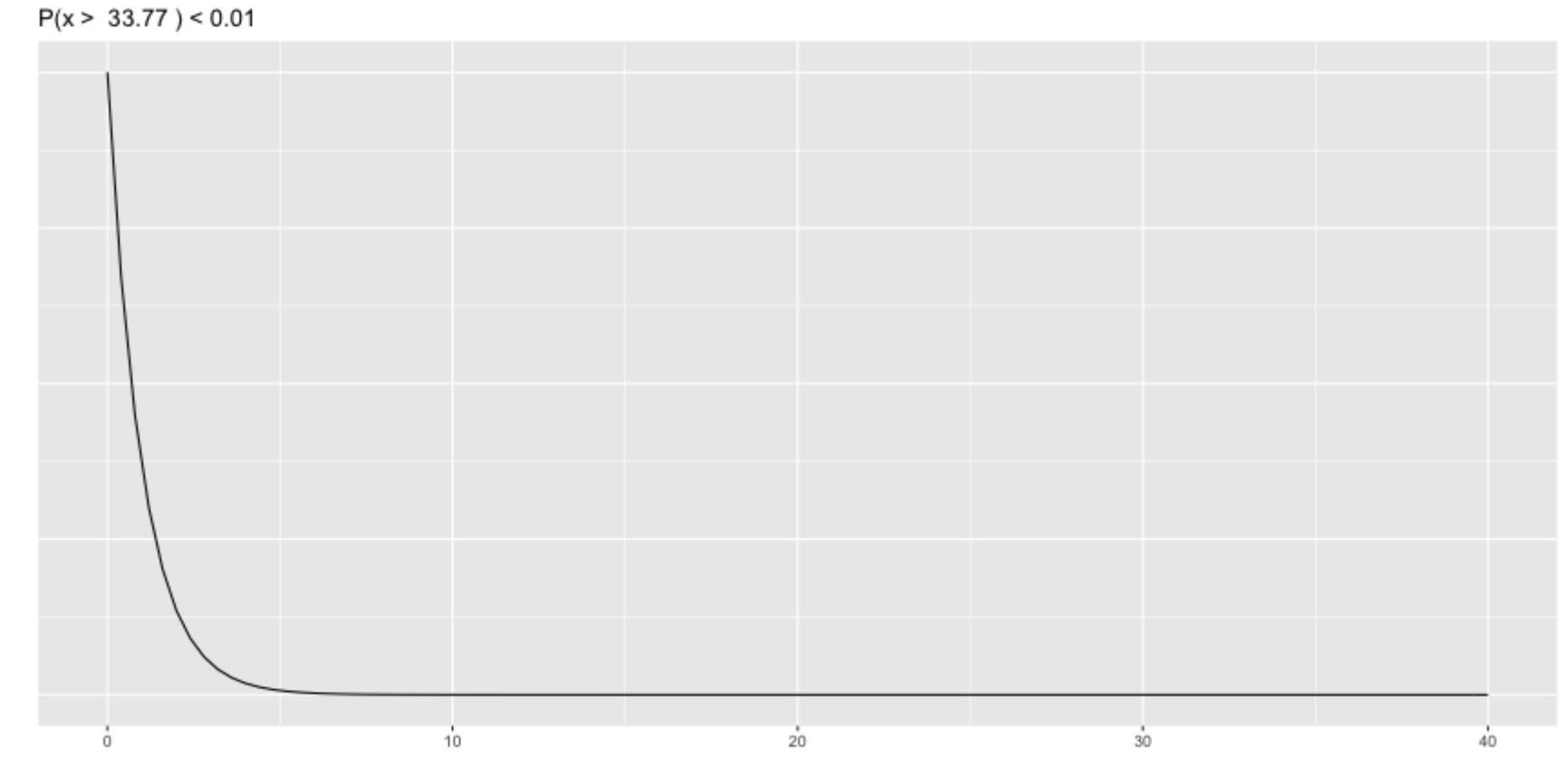
H_A: The mean GPA is different by ethnicities.

```
aov(GPA ~ Ethnicity, data = tutoring) %>% summary()
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Ethnicity       2   20.8  10.410   33.77 5.68e-15 ***
## Residuals    1139  351.2   0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis of Variance (ANOVA)

```
distribution_plot(stats::df, df1 = 2, df2 = 1139, cv = 33.77,
                  limits = c(0, 40), tails = 'greater')
```



Correlation

RQ: What is the relationship between age and GPA?

```
cor.test(tutoring$Age, tutoring$GPA)
```

```
##
##     Pearson's product-moment correlation
##
## data:  tutoring$Age and tutoring$GPA
## t = 1.1953, df = 1140, p-value = 0.2322
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02267598  0.09319808
## sample estimates:
##        cor
## 0.03537996
```

Linear Regression

RQ: Does age predict GPA?

```
lm.out <- lm(GPA ~ Age, data = tutoring)
summary(lm.out)
```

```
##
## Call:
## lm(formula = GPA ~ Age, data = tutoring)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2065 -0.2829  0.0492  0.3560  0.8717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.083683   0.071006  43.428   <2e-16 ***
## Age         0.002233   0.001868   1.195    0.232
## ---
```

Multiple Linear Regression

RQ: What student characteristics predict GPA?

```
lm.out <- lm(GPA ~ Gender + Ethnicity +
                   Military + ESL +
                   EdMother + EdFather +
                   Age + Employment +
                   Income + Transfer,
             data = tutoring)
```

```
##
## Call:
## lm(formula = GPA ~ Gender + Ethnicity + Military + ESL + EdMother +
##     EdFather + Age + Employment + Income + Transfer, data = tutoring)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.11193 -0.27820  0.02905  0.33182  1.34671
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.569086   0.120282  21.359  < 2e-16 ***
## GenderMALE      0.003155   0.038532   0.082  0.93476
## EthnicityOther  0.172046   0.056739   3.032  0.00248 **
## EthnicityWhite  0.312933   0.043814   7.142 1.64e-12 ***
## MilitaryTRUE   -0.074954   0.043941  -1.706  0.08832 .
## ESLTRUE        -0.055098   0.064869  -0.849  0.39585
## EdMother       -0.014198   0.012468  -1.139  0.25503
## EdFather        0.010899   0.011099   0.982  0.32633
## Age             0.001069   0.001977   0.541  0.58892
## Employment      0.039108   0.026088   1.499  0.13414
## Income          0.012751   0.007927   1.609  0.10799
## Transfer        0.003793   0.000701   5.410 7.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5481 on 1130 degrees of freedom
## Multiple R-squared:  0.08751,    Adjusted R-squared:  0.07863
## F-statistic: 9.852 on 11 and 1130 DF,  p-value: < 2.2e-16
```

Logistic Regression

RQ: What are the student characteristics that predict passing the course?

```
lr.out <- glm(Pass ~ treat2 + Gender + Ethnicity + Military + ESL +
              EdMother + EdFather + Age + Employment + Income + Transfer,
              data = tutoring,
              family = binomial(link = 'logit'))
```

Logistic Regression (cont.)

```
summary(lr.out)
```

```
##
## Call:
## glm(formula = Pass ~ treat2 + Gender + Ethnicity + Military +
##     ESL + EdMother + EdFather + Age + Employment + Income + Transfer,
##     family = binomial(link = "logit"), data = tutoring)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6086   0.2539   0.5273   0.6764   1.4138
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.682970   0.564704  -1.209   0.2265
## treat2TRUE      1.715245   0.316184   5.425 5.80e-08 ***
## GenderMALE     -0.354981   0.185785  -1.911   0.0560 .
## EthnicityOther  0.156269   0.247644   0.631   0.5280
## EthnicityWhite  0.791201   0.199122   3.973 7.08e-05 ***
## MilitaryTRUE    0.504942   0.214508   2.354   0.0186 *
## ESLTRUE        -0.119008   0.294073  -0.405   0.6857
## EdMother       -0.126442   0.059473  -2.126   0.0335 *
## EdFather        0.123463   0.055600   2.221   0.0264 *
## Age             0.002558   0.009728   0.263   0.7926
## Employment      0.253330   0.120726   2.098   0.0359 *
## Income          0.097734   0.039854   2.452   0.0142 *
## Transfer        0.005325   0.003492   1.525   0.1273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Questions

One Minute Paper

Complete the one minute paper:

https://forms.gle/yB3ds6MYE89Z1pURA

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?