Probability & Inference

Part 2: Multivariate PDFs and Hypothesis Testing

Today's Objectives

What if you have more than one variable, and you want to test whether they are associated?

- Multivariate PDFs
    - Joint Probability
    - Conditional Probability
    - Bayes Rule
- Two Variable (Bivariate) Hypothesis Tests
    - T-Tests
    - Chi-squared Tests
- Midterm Review

Multivariate PDFs

Multivariate PDFs

Now we have two variables: X and Y. Their **joint** probability distribution function must satisfy:

$$P(x, y) \geq 0$$

Discrete:

$$\sum_x \sum_y P(x, y) = 1$$

Continuous:

$$\int_x \int_y f(x, y) = 1$$

Example 1: Two Categorical Random Variables

```r
# Load CCES
CCES <- read_rds('data/CCES_2018.RDS') %>%
  mutate(gender = if_else(gender == 1, 'Male', 'Female'),
         age = 2018 - birthyr,
         party = case_when(pid3 == 1 ~ 'Democrat',
                           pid3 == 2 ~ 'Republican',
                           pid3 == 3 ~ 'Independent')) %>%
  filter(!is.na(party))

joint_distribution <- table(CCES$gender, CCES$party) / nrow(CCES)

joint_distribution
```

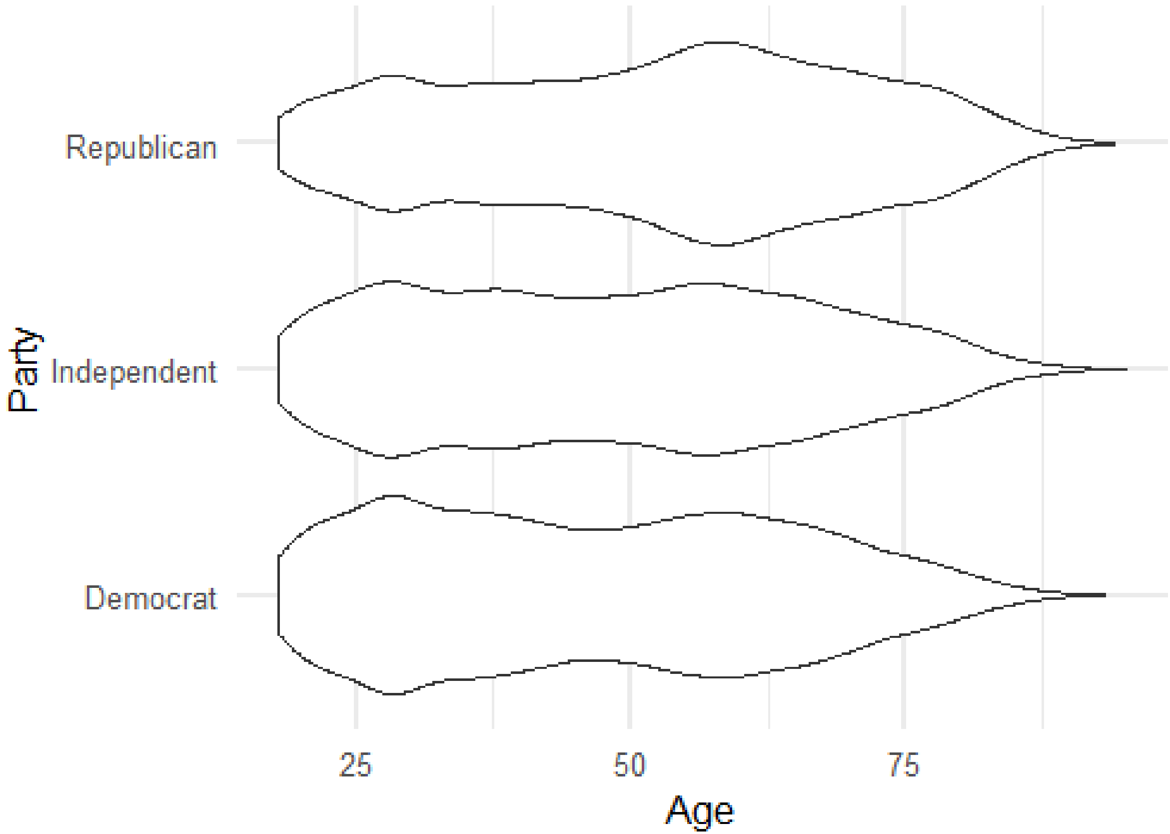
```
         Democrat Independent Republican
  Female 0.2536850   0.1532196  0.1581329
  Male   0.1427463   0.1548081  0.1374081
```

```r
sum(joint_distribution)
```

```
[1] 1
```

Example 2: One Categorical and One Continuous Random Variable

```r
ggplot(data = CCES) +
  geom_violin(mapping = aes(x=age, y = party)) +
  labs(x = 'Age', y = 'Party')
```



Marginal Distributions

The **marginal** distribution is the PDF one variable without considering the value of the other variable.

```r
joint_distribution # joint distribution of gender and party
```

```
         Democrat Independent Republican
  Female 0.2536850   0.1532196  0.1581329
  Male   0.1427463   0.1548081  0.1374081
```

```r
table(CCES$party) / nrow(CCES) # marginal distribution of party
```

```
   Democrat Independent  Republican
  0.3964313   0.3080276   0.2955410
```

```r
table(CCES$gender) / nrow(CCES) # marginal distribution of gender
```

```
   Female      Male
0.5650375 0.4349625
```

Marginal Distributions

Note: "Marginalizing" a distribution is equivalent to taking the row or column sums of the joint distribution.

```
table(CCES$party) / nrow(CCES)
```

```
   Democrat Independent  Republican 
  0.3964313   0.3080276   0.2955410 
```

```
colSums(joint_distribution) # marginal distribution of party
```

```
   Democrat Independent  Republican 
  0.3964313   0.3080276   0.2955410 
```

```
rowSums(joint_distribution)  # marginal distribution of gender
```

```
   Female      Male 
0.5650375 0.4349625 
```

Conditional Distributions

The **conditional** distribution is the PDF of one variable, holding the other variable constant.

```
joint_distribution
```

```
        Democrat Independent Republican
 Female 0.2536850   0.1532196  0.1581329
 Male   0.1427463   0.1548081  0.1374081
```

$$P(\text{party}|\text{gender}) = \frac{P(\text{party}, \text{gender})}{P(\text{gender})} = \frac{\text{joint}}{\text{marginal}}$$

```
# Conditional distribution of party given gender
joint_distribution / rowSums(joint_distribution)
```

```
        Democrat Independent Republican
 Female 0.4489703   0.2711670  0.2798627
 Male   0.3281807   0.3559113  0.3159079
```

Independence

Two variables are **independent** if the conditional distribution is the same as the marginal distribution.

$$P(\text{party}|\text{gender}) = P(\text{party})$$

Intuition: If men and women both have the same probability distribution over party, then we say that party is *independent* of gender.

Bayes Rule

$$P(\text{party}|\text{gender}) = \frac{P(\text{party}, \text{gender})}{P(\text{gender})}$$

and

$$P(\text{gender}|\text{party}) = \frac{P(\text{party}, \text{gender})}{P(\text{party})}$$

which means...

$$P(\text{gender}|\text{party})P(\text{party}) = P(\text{party}|\text{gender})P(\text{gender})$$

which means...

$$P(\text{gender}|\text{party}) = \frac{P(\text{party}|\text{gender})P(\text{gender})}{P(\text{party})}$$

Bayes Rule

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

If you know one conditional distribution, you can compute the other!

Bayes Rule

Suppose I get a positive COVID test. What's the chance I have COVID-19? I want to know $P(\text{COVID-negative}|\text{Positive Test})$.

I know the false positive rate of a COVID-19 test:

$$P(\text{Positive Test}|\text{COVID-negative}) = 0.05$$

I know my **prior** probability that I'm COVID-negative:

$$P(\text{COVID-negative}) = 0.95$$

I know the overall positivity rate in Georgia:

$$P(\text{Positive Test}) = 0.1$$

So, thanks to Bayes Rule, I know my **posterior** probability:

$$P(\text{COVID-negative}|\text{Positive Test}) = 0.05 \times \frac{0.95}{0.1} = 47.5\%$$

Bivariate Hypothesis Testing

Bivariate Hypothesis Testing

We have two variables and we want to know if they are **independent** of one another, or if there is an association.

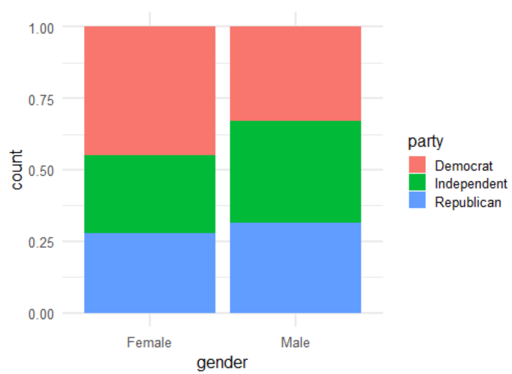
		Independent Variable	
		Categorical	Continuous
	Categorical	Tabular Analysis (chi-squared test)	MLE (probit/logit)
Dependent Variable	Continuous	Difference in Means (t-test)	OLS (linear regression)

Two Categorical Variables (Chi-Squared Test)

```
CCES %>%
  select(gender, party) %>%
  table
```

```
        party
gender   Democrat Independent Republican
  Female    13734        8295       8561
  Male       7728        8381       7439
```

```
ggplot(data = CCES) +
  geom_bar(mapping = aes(x=gender,fill=party), position = 'fill')
```



Chi-Squared Test

Step 1: Specify the Null Hypothesis

H_0: The two variables are **independent**.

Step 2: Generate the sampling distribution

Create a bunch of independent tables, and compute a chi-squared statistic for each.

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Step 3: Compare with observed outcome

Compare to the chi-squared statistic from the actual table to the sampling distribution.

Chi-Squared Test

Draw a sample and get the observed table.

```
n <- 1000

CCES_sample <- CCES %>%
  sample_n(size = n)

observed_table <- CCES_sample %>%
  select(gender, party) %>%
  table

observed_table
```

```
        party
gender   Democrat Independent Republican
  Female      270         151        161
  Male        142         138        138
```

Chi-Squared Test

What is the **expected** table if the two variables were independent?

```
gender_marginal_distribution <- table(CCES_sample$gender) / nrow(CCES
party_marginal_distribution <- table(CCES_sample$party) / nrow(CCES_s
expected_table <- outer(gender_marginal_distribution, party_marginal_
expected_table
```

```
        Democrat Independent Republican
 Female  239.784     168.198    174.018
 Male    172.216     120.802    124.982
```

Remember the definition of independence: conditional distributions are the same as the marginal distributions.

Chi-Squared Test

```r
get_null_chi_squared <- function(data, n){

  # get a random sample of the party variable
  party <- data %>%
    pull(party) %>%
    sample(size = n)

  # get a random sample of the gender variable
  gender <- data %>%
    pull(gender) %>%
    sample(size = n)

  # create the table
  null_table <- table(gender, party)

  # return the chi-squared statistic
  sum((null_table - expected_table)^2 / expected_table)
}

get_null_chi_squared(data = CCES_sample, n = 1000)
```

```
[1] 1.275859
```

Chi-Squared Test

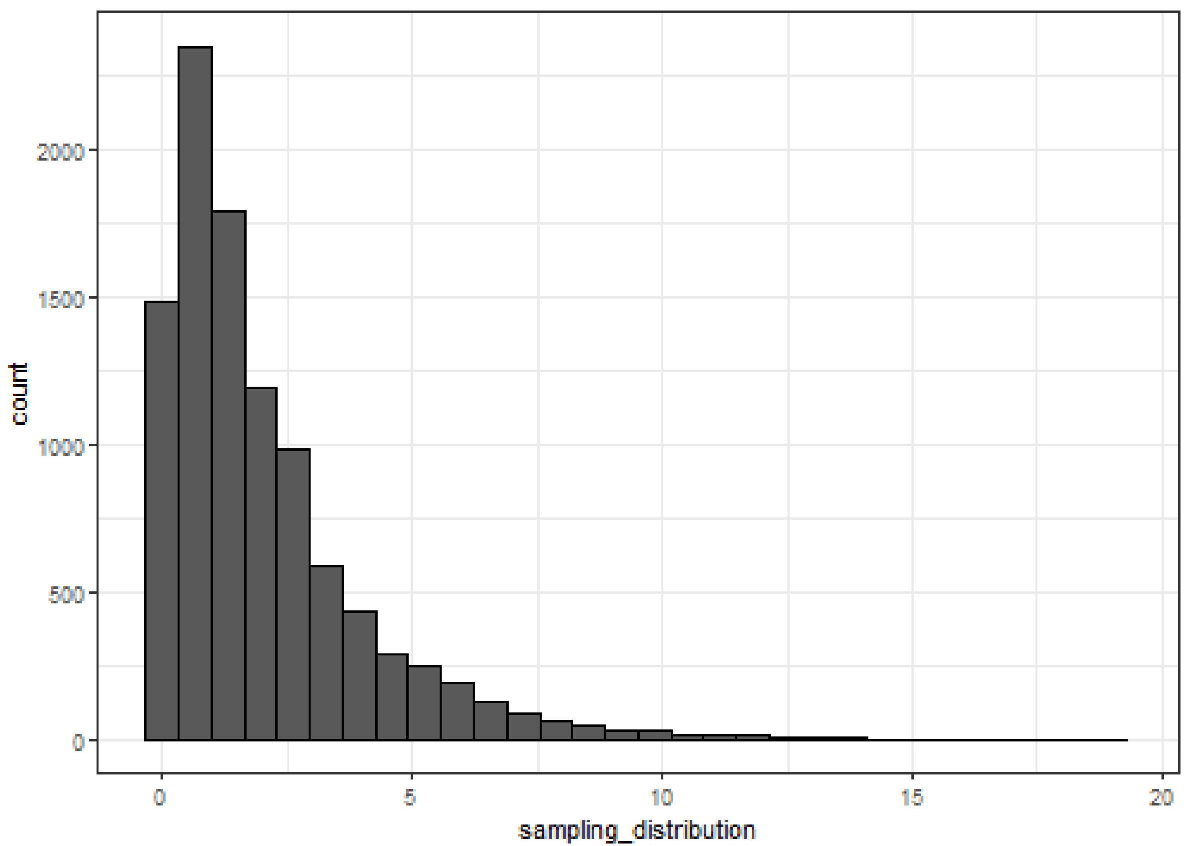
Generate the sampling distribution.

```
sampling_distribution <- replicate(10000, get_null_chi_squared(data =

chisq_plot <- tibble(sampling_distribution) %>%
  ggplot() +
  geom_histogram(aes(x=sampling_distribution), color = 'black') +
  theme_bw()
```

Chi-Squared Test

Plot the sampling distribution.

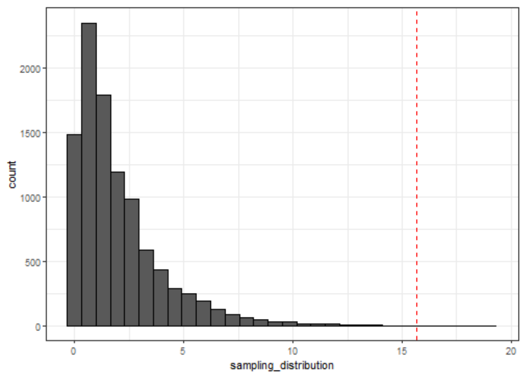
```
chisq_plot
```



Chi-Squared Test

Compare to the **actual** chi-squared statistic.

```
observed_chi_squared_statistic <- sum((observed_table - expected_tabl

chisq_plot +
  geom_vline(xintercept = observed_chi_squared_statistic, linetype =
```




```
# p-value
sum(sampling_distribution > observed_chi_squared_statistic) / length(
```

```
[1] 4e-04
```

Chi-Squared Test

Now, I can show you how to do it in one line…

```
CCES_sample %>%
  select(gender, party) %>%
  table
```

```
        party
gender   Democrat Independent Republican
  Female      270         151        161
  Male        142         138        138
```

```
CCES_sample %>%
  select(gender, party) %>%
  table %>%
  chisq.test
```

```

	Pearson's Chi-squared test

data:  .
X-squared = 15.646, df = 2, p-value = 0.0004005
```

One Categorical and One Continuous Variable (Two Sample T-Test)

One Categorical and One Continuous Variable (Two Sample T-Test)

Also known as a **difference in means** test.

```
CCES %>%
  group_by(party) %>%
  summarize(mean_age = mean(age))
```

```
# A tibble: 3 x 2
  party       mean_age
  <chr>          <dbl>
1 Democrat        46.6
2 Independent     48.0
3 Republican      52.2
```

Difference in Means Test

```r
# sample 1,000 Republicans ages
rep_age <- CCES %>%
  filter(party == 'Republican') %>%
  pull(age) %>%
  sample(100)

# sample 1,000 Democrats ages
dem_age <- CCES %>%
  filter(party == 'Democrat') %>%
  pull(age) %>%
  sample(100)

mean(rep_age)
```

```
[1] 53.91
```

```r
mean(dem_age)
```

```
[1] 44.33
```

The Republicans seem to be older on average, but is that just sampling error? How would you test it?

Difference in Means Test

Step 1: Specify the Null Hypothesis

H_0: There is no difference between the average age of Republicans and Democrats.

Step 2: Generate the Sampling Distribution

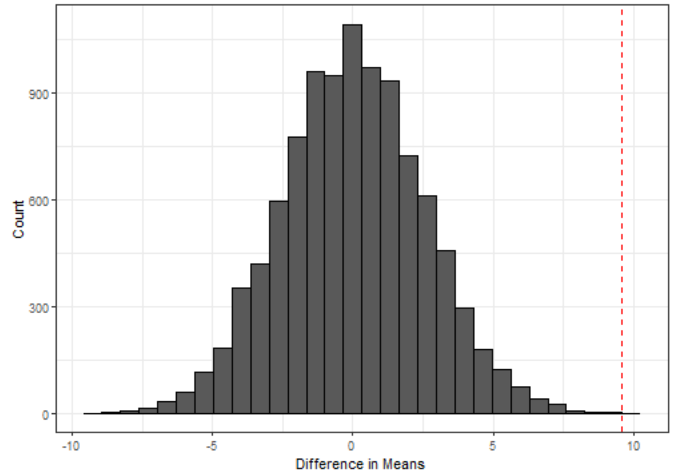
Function: Draw a Sample and Compute the Difference in Means

```r
difference_in_means <- function(population, n1 = 100, n2 = 100){

  # get the mean age of a random sample of size n1
  mean_age_dem <- population %>%
    pull(age) %>%
    sample(size = n1) %>%
    mean

  # get the mean age of a random sample of size n2
  mean_age_rep <- population %>%
    pull(age) %>%
    sample(size = n2) %>%
    mean

  # return the difference
  mean_age_rep - mean_age_dem
}

difference_in_means(CCES, n1 = 100, n2 = 100)
```

```
[1] 1.67
```

Step 2: Get the Sampling Distribution

```
observed <- mean(rep_age) - mean(dem_age)
sampling_distribution <- replicate(10000, difference_in_means(CCES, r

# sampling distribution
tibble(sampling_distribution) %>%
  ggplot() +
  geom_histogram(aes(x=sampling_distribution), color = 'black') +
  labs(x = 'Difference in Means', y = 'Count') +
  theme_bw() +
  geom_vline(xintercept = observed, linetype = 'dashed', color = 'rec
```



Step 3: Compare to Observed Test Statistic

```
# p-value
sum(abs(sampling_distribution) > observed) / length(sampling_distribu
```

```
[1] 1e-04
```

Difference in Means Test

Now I can show you how to do a two-sample t-test in one line…

```
t.test(rep_age, dem_age)
```

```

    Welch Two Sample t-test

data:  rep_age and dem_age
t = 4.036, df = 194.58, p-value = 7.812e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.898633 14.261367
sample estimates:
mean of x mean of y 
    53.91     44.33 
```

Difference in Means Test

Alternatively, you can use the "formula" syntax:

```
CCES %>%
  filter(party %in% c('Republican', 'Democrat')) %>%
  t.test(age ~ party, data = .)
```

```

	Welch Two Sample t-test

data:  age by party
t = -30.023, df = 34128, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.920804 -5.195107
sample estimates:
  mean in group Democrat mean in group Republican 
                46.64286                 52.20081 
```

Midterm Review