PSY 503: Foundations of Statistical Methods in Psychological Science

ANOVA, connections to lm()

Suyog Chandramouli

311 PSH (Princeton University)

19th November, 2025

Previously

- There is a commonality to statistical models
  - We are predicting an outcome
  - We have one or more measured variables

- Linear model
  - The form
    - $\text{outcome}_i = (\text{model}_i) + \text{error}_i$
    - $\widehat{\text{outcome}}_i = \hat{b}_0 + \hat{b}_1 \cdot \text{predictor}_i + \cdots + \text{error}_i$

  - Familiar models are a variant of a linear model

Previously

- What we've learned:
  - Y: an outcome
  - b_0: Intercept (baseline)
  - b's: Slopes (effects)
  - X's: Our predictors
  - ε: The stuff we can't explain (residuals)

$$\text{outcome}_i = (\text{model}_i) + \text{error}_i$$

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + \varepsilon$$

The Regression Family Tree

$$\text{outcome}_i = (\text{model}_i) + \text{error}_i$$

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + \varepsilon$$

- Multivariate Regression
  - Multiple Y's (MANOVA, etc.)
- Univariate Regression
  - Multiple regression: $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + \varepsilon$
    - Multiple predictors
    - Each b tells us about unique contribution
  - Simple linear regression: $Y = b_0 + b_1X_1 + \varepsilon$
    - Single predictor
    - b_1 is related to Pearson's correlation (r)
  - Simplest model: $Y = b_0 + \varepsilon$
    - Just the intercept
    - Average with deviation

- Questions we can ask:
  - Is b_0 different from some value?
  - Is b_1 different from 0?
  - How many predictors do we really need?

Comparing Two Means

Z-test

- You know these:
  - population mean – μ_0
  - population SD – σ
  - You have a sample
    - mean – $\bar{X}$
    - Sample size - N

- Question: Is $\bar{X}$ different from μ_0?

Z-test

You know these:

- population mean – μ_0
- population SD – σ
- You have a sample
  - mean – $\bar{X}$
  - Sample size - N

Question: Is $\bar{X}$ different from μ_0?

Remember z-score?

$$z = (X - \mu) / \sigma$$

Standardized a single observation

Test-statistic

- Deals with a sample

$$z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$$

$$z = (\bar{X} - \mu) / SE$$

- "How many SEs is this sample from expected?"

Z-test

You know these:

- population mean – μ_0
- population SD – σ
- You have a sample
  - mean – $\bar{X}$
  - Sample size - N

Question: Is $\bar{X}$ different from μ_0?

Remember z-score?

$$z = (X - \mu) / \sigma$$

Standardized a single observation

Test-statistic

- Deals with a sample

$$z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$$

$$z = (\bar{X} - \mu) / SE$$

Observed difference / Typical variability

- "How many SEs is this sample from expected?"

Z-test

"one of the most useless tests in all of statistics: the z-test. Seriously – this test is almost never used in real life." – Danielle Navarro

T-tests

- Three kinds of t-tests
  - One-sample
  - Two-sample (independent sample)
  - Paired-sample

T-tests

- Three kinds of t-tests
  - One-sample - compare one group to a constant
  - Two-sample (***independent*** sample) - compare two ***independent*** groups
  - Paired-sample - compare matched observations

One-sample t-test

- Do we differ from a known value?
    - Is the average height of the class different from 6ft?

One sample t-test

You know these:

- population mean – μ_0
- You have a sample
  - mean – $\bar{X}$
  - Sample size – N
  - Sample SD – **s**

Question: **Is $\bar{X}$ different from μ_0?**

Test-statistic

- Earlier

  $z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

  $z = (\bar{X} - \mu) / SE$

- Now, we are estimating σ
  - $t = (\bar{X} - \mu) / (s/\sqrt{N})$

- "How many SEs is this sample from expected?"

One sample t-test

You know these:

- population mean – μ_0
- You have a sample
  - mean – $\bar{X}$
  - Sample size – N
  - Sample SD – **s**

Question: **Is $\bar{X}$ different from μ_0?**

Test-statistic

- Earlier

  $z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

  $z = (\bar{X} - \mu) / SE$

- Now, we are estimating σ
  - $t = (\bar{X} - \mu) / (s/\sqrt{N})$

- "How many SEs is this sample from expected?"

Observed difference /
***Estimated** Typical variability*

One-sample t-test

- Do we differ from a known value?
  - Is the average height of the class different from 6ft?

z-test vs t-test

z-test: know population **SD**
t-test: estimate **SD** from sample

One-sample t-test

- Do we differ from a known value?
  - Is the average height of the class different from 6ft?

z-test vs t-test

z-test: know population **SD**
t-test: estimate **SD** from sample

CLT guarantees

- Means are normal.
- Proportions are normal (large N).

One-sample t-test

- Do we differ from a known value?
  - Is the average height of the class different from 6ft?

z-test vs t-test

z-test: know population **SD**
t-test: estimate **SD** from sample

CLT guarantees

- Means are normal.
- Proportions are normal (large N).

test-statistic, sampling distributions

z-test: know population **SD**

$z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

→ use normal

t-test: estimate **SD** from sample

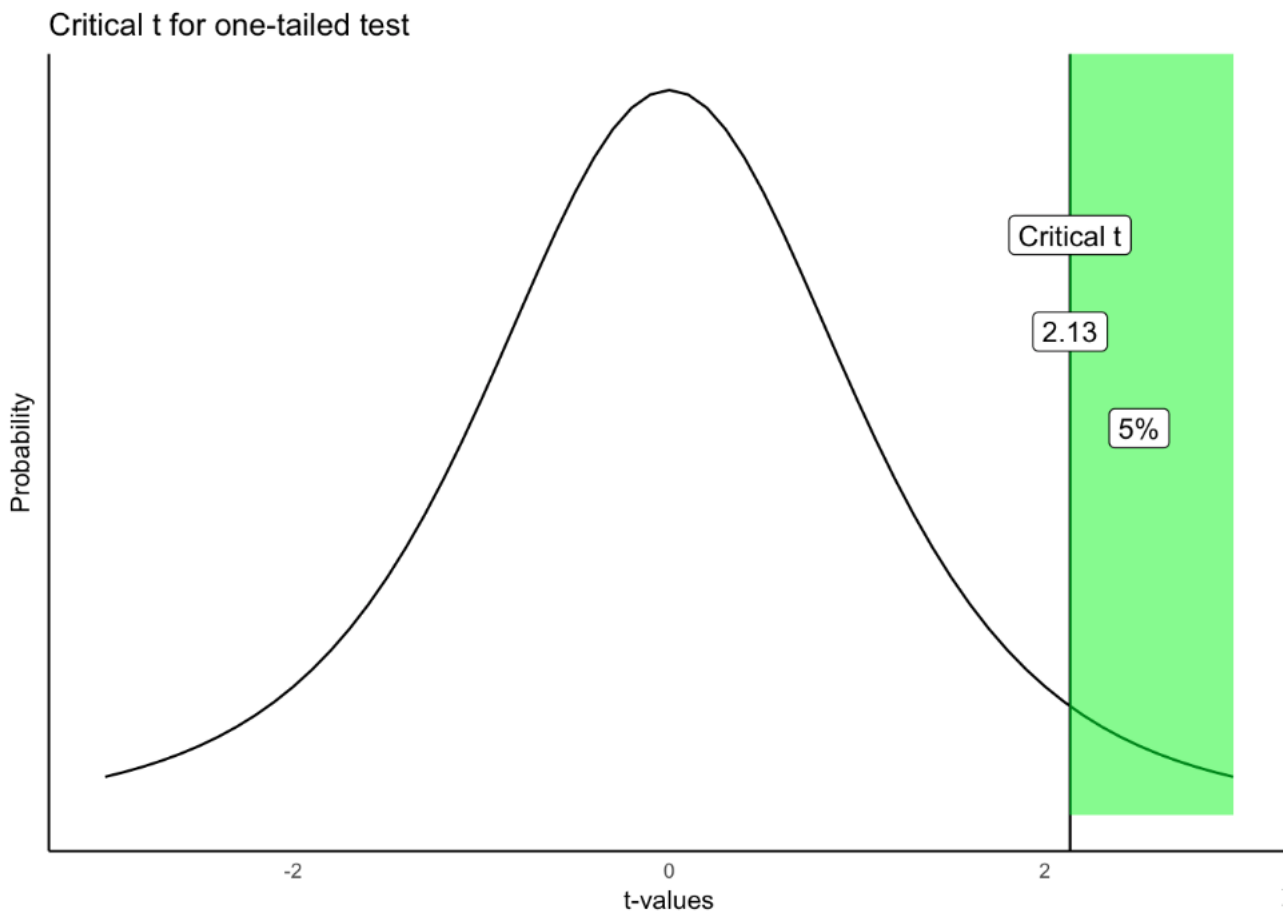
$t = (\bar{X} - \mu) / (\mathbf{s}/\sqrt{(N-1)})$

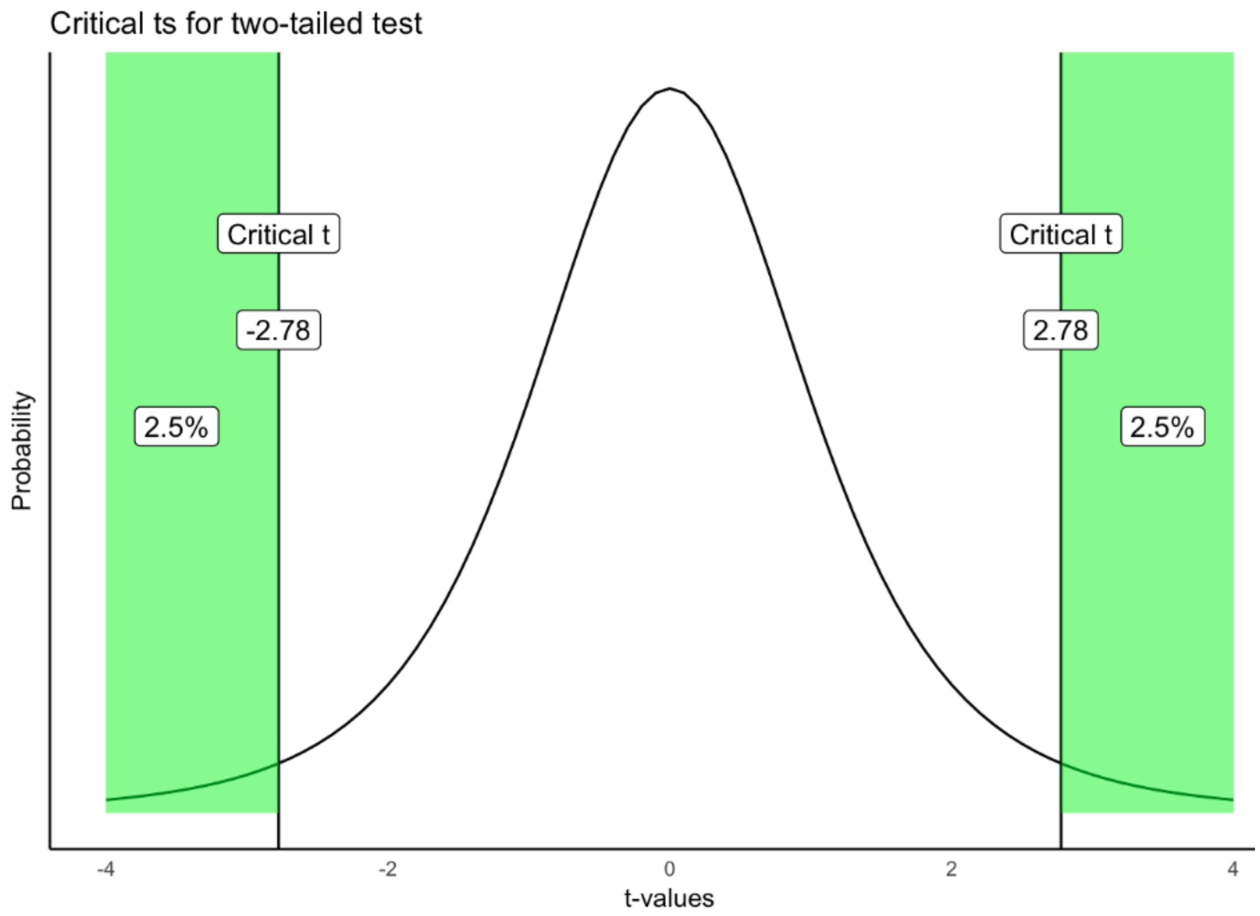
→ use t-distribution

T-distribution vs Gaussian

- A Shiny demo: https://shiny.abdn.ac.uk/Stats/apps/app_distn/

- Key differences
  - T-distribution has heavier tails
    - When n (or df) is small, the tail is heavier
    - As n increases, t-distribution resembles the Gaussian more
  - Implications?


Critical t for one-tailed test
Critical t
2.13
5%
Probability
-2
0
2
t-values


Critical ts for two-tailed test
Critical t
-2.78
2.5%
Critical t
2.78
2.5%
Probability
-4
-2
0
2
4
t-values

Sampling distributions and connections to Normal

- χ^2 – sum of squared normals.
- t – ratio of normal and $\sqrt{\chi^2}$
- F – ratio of two χ^2.

```{r}
set.seed(123)
data <- rnorm(30, mean=1, sd=2)

# One-sample t-test
t_model <- t.test(data, mu=0)
print(t_model)
```

```
        One Sample t-test

data:  data
t = 2.5286, df = 29, p-value = 0.01715
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1731467 1.6384383
sample estimates:
mean of x
0.9057925
```

t-test and linear model

- **Use Case:** Comparing a single group's mean to a known value/reference point

- **As a linear model:**
  - $Y = b_0 + \varepsilon$
    - *Model without predictors*

  - NHST
    - Traditional form: $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu$
    - lm() equivalent: $H_0: b_0 = \mu_0$ vs $H_1: b_0 \neq \mu_0$

But where is the t-statistic in lm()?

- Linear Model
  - $y = \beta_0 + \varepsilon$
  - Estimate: $b_0 = \bar{x}$
  - Standard Error: $SE(b_0) = s/\sqrt{n}$
  - t-statistic: $(b_0 - \beta_0)/(s/\sqrt{n})$
- T-test
  - β_0 in linear model = μ in t-test
  - Both estimate central tendency
  - Both use same standard error formula

```{r}
set.seed(123)
data <- rnorm(30, mean=1, sd=2)

# One-sample t-test
t_model <- t.test(data, mu=0)
print(t_model)
```

```
        One Sample t-test

data:  data
t = 2.5286, df = 29, p-value = 0.01715
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1731467 1.6384383
sample estimates:
mean of x
0.9057925
```

```{r}
lm_model<- lm(data ~ 1)  # intercept-only model
summary(lm_model)
```

```
Call:
lm(formula = data ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8390 -1.2486 -0.0533  1.0714  3.6680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9058     0.3582   2.529   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.962 on 29 degrees of freedom
```

````
```{r}
set.seed(123)
data <- rnorm(30, mean=1, sd=2)

# One-sample t-test
t_model <- t.test(data, mu=0)
print(t_model)
```
````

```
        One Sample t-test

data:  data
t = 2.5286, df = 29, p-value = 0.01715
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1731467 1.6384383
sample estimates:
mean of x
0.9057925
```

````
```{r}
lm_model<- lm(data ~ 1)  # intercept-only model
summary(lm_model)
```
````

```
Call:
lm(formula = data ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8390 -1.2486 -0.0533  1.0714  3.6680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9058     0.3582   2.529   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.962 on 29 degrees of freedom
```

```{r}
set.seed(123)
data <- rnorm(30, mean=1, sd=2)

# One-sample t-test
t_model <- t.test(data, mu=0)
print(t_model)
```

```
        One Sample t-test

data:  data
t = 2.5286, df = 29, p-value = 0.01715
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1731467 1.6384383
sample estimates:
mean of x
0.9057925
```

```{r}
lm_model<- lm(data ~ 1)  # intercept-only model
summary(lm_model)
```

```
Call:
lm(formula = data ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8390 -1.2486 -0.0533  1.0714  3.6680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9058     0.3582   2.529   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.962 on 29 degrees of freedom
```

```{r}
set.seed(123)
data <- rnorm(30, mean=1, sd=2)

# One-sample t-test
t_model <- t.test(data, mu=0)
print(t_model)
```

```
        One Sample t-test

data:  data
t = 2.5286, df = 29, p-value = 0.01715
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1731467 1.6384383
sample estimates:
mean of x
0.9057925
```

```{r}
lm_model<- lm(data ~ 1)  # intercept-only model
summary(lm_model)
```

```
Call:
lm(formula = data ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8390 -1.2486 -0.0533  1.0714  3.6680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9058     0.3582   2.529   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.962 on 29 degrees of freedom
```

```{r}
set.seed(123)
data <- rnorm(30, mean=1, sd=2)

# One-sample t-test
t_model <- t.test(data, mu=0)
print(t_model)
```

```
        One Sample t-test

data:  data
t = 2.5286, df = 29, p-value = 0.01715
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1731467 1.6384383
sample estimates:
mean of x
0.9057925
```

```{r}
lm_model<- lm(data ~ 1)  # intercept-only model
summary(lm_model)
```

```
Call:
lm(formula = data ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8390 -1.2486 -0.0533  1.0714  3.6680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9058     0.3582   2.529   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.962 on 29 degrees of freedom
```

Two-sample t-test
(Independent samples t-test)

- Do two groups differ (between-subjects design)?
  - Treatment vs. control
    - Did the manipulation (A vs. B) cause a difference in the measure?

subjects_A	A	subjects_B	B
1	1	6	4
2	4	7	8
3	3	8	7
4	6	9	9
5	5	10	10

Two sample t-test

You know these:

- You have sample from Group 1
  - mean – $\bar{X}_1$
  - Sample size – N_1
  - Sample SD – s_1
- You have sample from Group 2
  - mean – $\bar{X}_2$
  - Sample size – N_2
  - Sample SD – s_2

Question: Is $\bar{X}_1$ different from $\bar{X}_2$?
(Are the two groups different?)

Test-statistic

- Earlier

  $z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

- Now,
  - $t = (\bar{X}_1 - \bar{X}_2) / SE\_diff$
    - $SE\_diff = \sqrt{(s_1^2/N_1 + s_2^2/N_2)}$
- "How many SEs apart are these two samples?"

Two sample t-test

You know these:

- You have sample from Group 1
  - mean – $\bar{X}_1$
  - Sample size – N_1
  - Sample SD – s_1

- You have sample from Group 2
  - mean – $\bar{X}_2$
  - Sample size – N_2
  - Sample SD – s_2

Question: Is $\bar{X}_1$ different from $\bar{X}_2$?
(Are the two groups different?)

Test-statistic

- Earlier

  $z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

- Now,
  - $t = (\bar{X}_1 - \bar{X}_2) / SE\_diff$
    - $SE\_diff = \sqrt{(s_1^2/N_1 + s_2^2/N_2)}$
- "How many SEs is this sample from expected?"

Observed difference /
***Estimated** Typical variability*

Two sample t-test

You know these:

- You have sample from Group 1
  - mean – $\bar{X}_1$
  - Sample size – N_1
  - Sample SD – s_1
- You have sample from Group 2
  - mean – $\bar{X}_2$
  - Sample size – N_2
  - Sample SD – s_2

Question: Is $\bar{X}_1$ different from $\bar{X}_2$?
(Are the two groups different?)

Test-statistic

- Earlier

  $z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

- Now,
  - $t = (\bar{X}_1 - \bar{X}_2) / SE\_diff$
    - $SE\_diff = \sqrt{(s_1^2/N_1 + s_2^2/N_2)}$
- "How many SEs is this sample from expected?"

Df depends on assumptions:
$\sigma_1^2 = \sigma_2^2$ (equal variances)
$\Rightarrow$ *Pooled t-test*

$\sigma_1^2 != \sigma_2^2$ (equal variances)
$\Rightarrow$ *Welch's t-test*

Observed difference /
Estimated *Typical variability*

Pooled t-test vs Welch's t-test

- Pooled t-test
  - **Assumption**
    $\sigma_1^2 = \sigma_2^2$ (equal variances)
  - **Estimation**
    Pool both samples to estimate one common variance
    - df = $N_1 + N_2 - 2$
    - SE_diff = s_pooled × $\sqrt{(1/N_1 + 1/N_2)}$
- Welch's t-test
  - **Assumption**
    $\sigma_1^2 \mathrel{!}= \sigma_2^2$ (unequal variances)
  - **Estimation**
    Use separate variance estimates
    - df = complex formula (Welch-Satterthwaite)
      - Often not an integer
      - df is always ≤ $N_1 + N_2 - 2$
      - More conservative when variances differ

Pooled t-test vs Welch's t-test

- Pooled t-test
  - **Assumption**
    $\sigma_1^2 = \sigma_2^2$ (equal variances)
  - **Estimation**
    Pool both samples to estimate one common variance
    - df = $N_1 + N_2 - 2$
    - SE_diff = s_pooled × $\sqrt{(1/N_1 + 1/N_2)}$
- Welch's t-test
  - **Assumption**
    $\sigma_1^2 \mathrel{!}= \sigma_2^2$ (unequal variances)
  - **Estimation**
    Use separate variance estimates
    - df = complex formula (Welch-Satterthwaite)
      - Often not an integer
      - df is always ≤ $N_1 + N_2 - 2$
      - More conservative when variances differ

Pro Tip: Use Welch's by default. It's safer and performs well even when variances are equal.

Fyi, df for Welch's t-test

$$df = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}}$$

Two-sample t-test

- **Use Case:** Comparing means between two independent groups

- **As a Linear Model:**
  - $Y = b_0 + b_1X_1 + \varepsilon$
    - where X_1 is dummy coded (0,1)

  - NHST
    - Traditional Form: $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$
    - lm() equivalent: $H_0: b_1 = 0$ vs $H_1: b_1 \neq 0$

Two-sample: t.test () vs lm()

```
#Careful of defaults!

#default: Welch's
t.test(Y ~ X1, data = df)

#Change default to pool!
t.test(Y ~ X, var.equal = TRUE)
```

Two-sample: t.test () vs lm()

```
#Careful of defaults!

#default: Welch's
t.test(Y ~ X1, data = df)

#Change default to pool!
t.test(Y ~ X, var.equal = TRUE)
```

```
#default: Pooled
lm(Y ~ X1, data = df)
#NOTE: lm() not equipped for welch's
```

From: https://steverxd.github.io/Stat_tests/two-means-independent-samples.html

Two-sample: t.test () vs lm()

```r
#Careful of defaults!

#default: Welch's
t.test(Y ~ X1, data = df)

#Change default to pool!
t.test(Y ~ X, var.equal = TRUE)
```

```r
#default: Pooled
lm(Y ~ X1, data = df)
#NOTE: lm() not equipped for welch's
```

```r
# Linear model (GLS)
lm <- nlme::gls(value ~ 1 + group_y2, weights = nlme::varIdent(form=~1|group), method="ML"
  lm %>% summary() %>% print(digits = 8) # show summary output
  confint(lm) # show confidence intervals
```

From: https://steverxd.github.io/Stat_tests/two-means-independent-samples.html

Paired Sample t-test

- Do two samples differ (within-subjects design)?
  - Same participants
    - Compare
      - Before Treatment vs. After Treatment

subjects	level_A	level_B
1	1	4
2	4	8
3	3	7
4	6	9
5	5	10

Paired sample t-test

You know these:

- You have paired observations
  - Before/after for each subject
  - Or matched pairs
- From this you can construct a difference distribution/ sample
  - $D = X\_after - X\_before$
    - mean difference – $\bar{D}$
    - SD of differences – s_D
  - Sample size – N

Question: Is $\bar{D}$ different from 0? (Did the treatment?)

Test-statistic

- Earlier

$$z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$$
$$t = (\bar{X} - \mu) / (s/\sqrt{N})$$
$$t = (\bar{X}_1 - \bar{X}_2) / SE\_diff$$

- Now, treats differences as one sample
  - $t = \bar{D} / (s\_D/\sqrt{N})$ • $t = \bar{D} / SE\_D$
- " How many SEs is the mean difference from zero?"

Paired sample t-test

You know these:

- You have paired observations
  - Before/after for each subject
  - Or matched pairs
- From this you can construct a difference distribution/ sample
  - $D = X\_after - X\_before$
    - mean difference – $\bar{D}$
    - SD of differences – s_D
  - Sample size – N

Question: Is $\bar{D}$ different from 0? (Did the treatment?)

Test-statistic

- Earlier

  $z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$
  $t = (\bar{X} - \mu) / (s/\sqrt{N})$
  $t = (\bar{X}_1 - \bar{X}_2) / SE\_diff$

- Now, treats differences as one sample
  - $t = \bar{D} / (s\_D/\sqrt{N})$ • $t = \bar{D} / SE\_D$
- " How many SEs is the mean difference from zero?"

Observed difference /
***Estimated** Typical variability*

Paired t-test

- **Use Case:** Comparing means between two related/matched measurements (e.g., before vs. after)

- **As a Linear Model:**
  - $D = b_0 + \varepsilon$
    - where $D = Y_2 - Y_1$ (differences)

  - NHST
    - Traditional Form: H_0: $\mu D = 0$ vs H_1: $\mu D \neq 0$
    - lm() equivalent: H_0: $b_0 = 0$ vs H_1: $b_0 \neq 0$

Paired t-test – t.test() vs lm()

```r
#handles pairing automatically
t.test(after, before, paired = TRUE)
```

```r
# lm works with the difference distribution
D <- after - before
lm(D ~ 1)  # intercept-only
```

Non-parametric t-tests

- Parametric tests still make some strong assumptions

- Non-parametric tests can be desirable when
  - Data is not continuous
  - Skewed generating (population) distribution
  - Heavy outliers

- Independent groups: **Mann-Whitney U test**
- Paired samples: **Wilcoxon signed-rank test**.

Non-parametric t-tests

- Previously
  - **Bootstrapping:**
    - Sample *with replacement*. Estimates sampling variability.
- Similarly,
  - **Permutation (for Mann-Whitney/Wilcoxon):**
    - Shuffle group labels. Tests null hypothesis directly.

Mann-Whitney algorithm

Sampling distribution:

1. Calculate observed U statistic
2. Randomly shuffle group labels
3. Recalculate U
4. Repeat 10,000+ times

Now, where does observed U fall?

Mann-Whitney algorithm

Sampling distribution:

1. Calculate observed U statistic
2. Randomly shuffle group labels
3. Recalculate U
4. Repeat 10,000+ times

Now, where does observed U fall?

- **U statistic:**
  - Count pairwise comparisons.
  - For each observation in Group 1, count how many in Group 2 it beats.
  - Sum those counts = U
  - Example:
    - Group 1: [3, 7]
    - Group 2: [1, 4, 5]
      - 3 beats 1 → +1
        3 loses to 4, 5 → +0
      - 7 beats 1, 4, 5 → +3
      - U = 4

Wilcoxon signed-rank via permutation

Sampling distribution:

1. Calculate signed ranks of differences
2. Randomly flip signs of differences
3. Recalculate test statistic
4. Repeat

Now, where does observed stat fall?

Wilcoxon signed-rank via permutation

Sampling distribution:

1. Calculate signed ranks of differences
2. Randomly flip signs of differences
3. Recalculate test statistic
4. Repeat

Now, where does observed stat fall?

- **Signed ranks:**
  - Take differences.
  - Ignore sign, rank by magnitude.
  - Put signs back.
  - Example:
    - Differences: [+5, -2, +8, -1]
    - Absolute: [5, 2, 8, 1]
    - Ranks: [3, 2, 4, 1]
    - Signed ranks: [+3, -2, +4, -1]
    - Sum positive ranks.
    - Sum negative ranks.
    - Compare.

Comparing Several Means

One-way ANOVA

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

One-way ANOVA

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

- Not: "Which ones differ?”
- Not: "By how much?"
- Just: "Is there a difference *somewhere*?"

One-way ANOVA

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

- Not: "Which ones differ?"
- Not: "By how much?"
- Just: "Is there a difference *somewhere*?"

Test-statistic

- Earlier

  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

Now, comparing k means

$$F = \frac{MS_{between}}{MS_{within}}$$

"How much do groups vary compared to typical within-group noise?"

One-way ANOVA

Between group variation / ***variation within groups***

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

- Not: "Which ones differ?"
- Not: "By how much?"
- Just: "Is there a difference *somewhere*?"

Test-statistic

- Earlier

  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

Now, comparing k means

$$F = \frac{MS_{between}}{MS_{within}}$$

"How much do groups vary compared to typical within-group noise?"

One-way ANOVA

Variation between group means /
Remaining unexplained variance

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

- Not: "Which ones differ?”
- Not: "By how much?"
- Just: "Is there a difference *somewhere*?"

Test-statistic

- Earlier
  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

Now, comparing k means

$$F = \frac{MS_{between}}{MS_{within}}$$

“How much do groups vary compared to typical within-group noise?”

One-way ANOVA

Variance due to manipulation / random error

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

- Not: "Which ones differ?"
- Not: "By how much?"
- Just: "Is there a difference *somewhere*?"

Test-statistic

- Earlier

  $t = (\bar{X} - \mu) / (s/\sqrt{N})$

Now, comparing k means

$$F = \frac{MS_{between}}{MS_{within}}$$

"How much do groups vary compared to typical within-group noise?"

Observed F is computed directly from the data:

Source	**df**	**SS**
Effect	$k - 1$	SS_{Effect}
Error	$n - k$	SS_{Error}

k = number of groups; n = total sample-size

$$SS_{Effect} = \sum_{i=1}^{k} n_i (X_i - \bar{X})^2$$

$$SS_{Error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)^2$$

Observed F is computed directly from the data:

Source	df	SS	MSE
Effect	$k-1$	SS_{Effect}	$MS_{Effect} = \frac{SS_{Effect}}{k-1}$
Error	$n-k$	SS_{Error}	$MS_{Error} = \frac{SS_{Error}}{n-k}$

k = number of groups; n = total sample-size

$$SS_{Effect} = \sum_{i=1}^{k} n_i (X_i - \bar{X})^2$$

$$SS_{Error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)^2$$

Observed F is computed directly from the data:

Source	df	SS	MSE	F	p
Effect	$k-1$	SS_{Effect}	$MS_{Effect} = \frac{SS_{Effect}}{k-1}$	$\frac{MS_{Effect}}{MS_{Error}}$	Calculated from F-distribution
Error	$n-k$	SS_{Error}	$MS_{Error} = \frac{SS_{Error}}{n-k}$		

k = number of groups; n = total sample-size

$$SS_{Effect} = \sum_{i=1}^{k} n_i (X_i - \bar{X})^2$$

$$SS_{Error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)^2$$

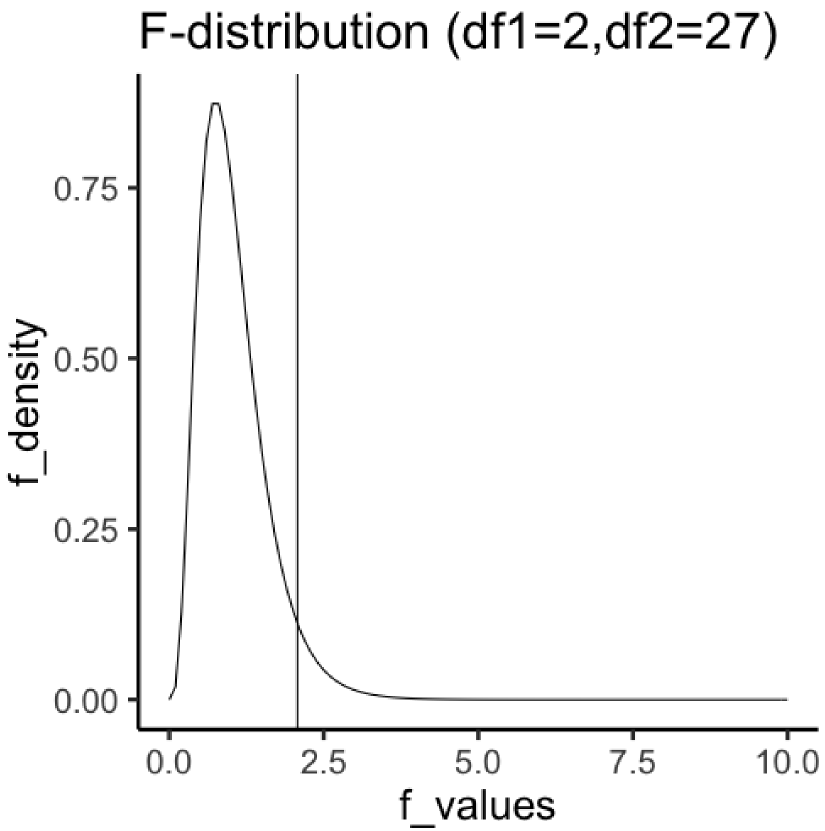
One-way ANOVA

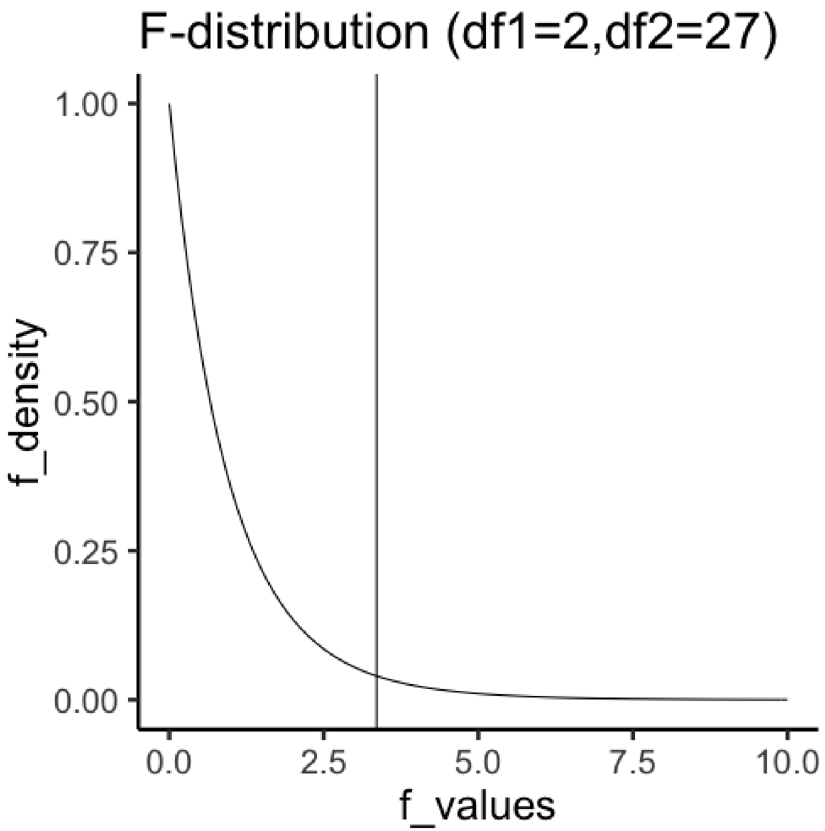
- **Test Statistic**

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

- **Global test**
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  - Our F-test tells us if we can reject this

Example:
ANOVA: H_0: $\mu_1 = \mu_2 = \mu_3$
↓ (Reject H_0)


F-distribution (df1=2,df2=27)
f_density
f_values
0.00
0.25
0.50
0.75
0.0
2.5
5.0
7.5
10.0


F-distribution (df1=2,df2=27)
f_density
f_values
0.00
0.25
0.50
0.75
1.00
0.0
2.5
5.0
7.5
10.0

Example

````
```{r}
A <- c(20,11,2)
B <- c(6,2,7)
C <- c(2,11,2)
IV <- as.factor(rep(c("A","B","C"),each=3))
DV <- c(A,B,C)
df <- data.frame(IV,DV)
```
````

Example

```{r}
A <- c(20,11,2)
B <- c(6,2,7)
C <- c(2,11,2)
IV <- as.factor(rep(c("A","B","C"),each=3))
DV <- c(A,B,C)
df <- data.frame(IV,DV)
```

```{r}
df
```

IV <fctr>	DV <dbl>
A	20
A	11
A	2
B	6
B	2
B	7
C	2
C	11
C	2

9 rows

Example

```{r}
A <- c(20,11,2)
B <- c(6,2,7)
C <- c(2,11,2)
IV <- as.factor(rep(c("A","B","C"),each=3))
DV <- c(A,B,C)
df <- data.frame(IV,DV)
```

```{r}
df
```

IV <fctr>	DV <dbl>
A	20
A	11
A	2
B	6
B	2
B	7
C	2
C	11
C	2

9 rows

```{r}
aov(DV~IV,df)
```

```
Call:
   aov(formula = DV ~ IV, data = df)

Terms:
                IV Residuals
Sum of Squares  72       230
Deg. of Freedom  2         6

Residual standard error: 6.191392
Estimated effects may be unbalanced
```

Example

```{r}
A <- c(20,11,2)
B <- c(6,2,7)
C <- c(2,11,2)
IV <- as.factor(rep(c("A","B","C"),each=3))
DV <- c(A,B,C)
df <- data.frame(IV,DV)
```

```{r}
df
```

IV <fctr>	DV <dbl>
A	20
A	11
A	2
B	6
B	2
B	7
C	2
C	11
C	2

9 rows

```{r}
aov(DV~IV,df)
```

```
Call:
   aov(formula = DV ~ IV, data = df)

Terms:
                IV Residuals
Sum of Squares  72       230
Deg. of Freedom  2         6

Residual standard error: 6.191392
Estimated effects may be unbalanced
```

```{r}
aov_results <- aov(DV~IV,df)
summary(aov_results)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
IV           2     72   36.00   0.939  0.442
Residuals    6    230   38.33
```

Example

```{r}
A <- c(20,11,2)
B <- c(6,2,7)
C <- c(2,11,2)
IV <- as.factor(rep(c("A","B","C"),each=3))
DV <- c(A,B,C)
df <- data.frame(IV,DV)
```

```{r}
df
```

IV <fctr>	DV <dbl>
A	20
A	11
A	2
B	6
B	2
B	7
C	2
C	11
C	2

9 rows

```{r}
aov(DV~IV,df)
```

```
Call:
   aov(formula = DV ~ IV, data = df)

Terms:
                 IV Residuals
Sum of Squares   72       230
Deg. of Freedom   2         6

Residual standard error: 6.191392
Estimated effects may be unbalanced
```

```{r}
aov_results <- aov(DV~IV,df)
summary(aov_results)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
IV           2     72   36.00   0.939  0.442
Residuals    6    230   38.33
```

F(2,6) = 0.939, p = 0.442

One-way ANOVA

- **Use Case:** Comparing means across three or more independent groups

- **As a Linear Model:**
  - $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + \varepsilon$
    - where X's are dummy coded for k-1 groups

  - NHST
    - Traditional Form: $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ vs H_1: not all μ_i equal
    - lm() equivalent: $H_0: b_1 = b_2 = \ldots = 0$ vs H_1: not all $b_i = 0$

Try it

- What do you expect F to be when all groups have the same μ’s?
  - What do you observe?
  - Now separate them. What happens to F? What happens to p?
  - When is it negative?
- Is the eyeball test generally valid? What do you think? Try?

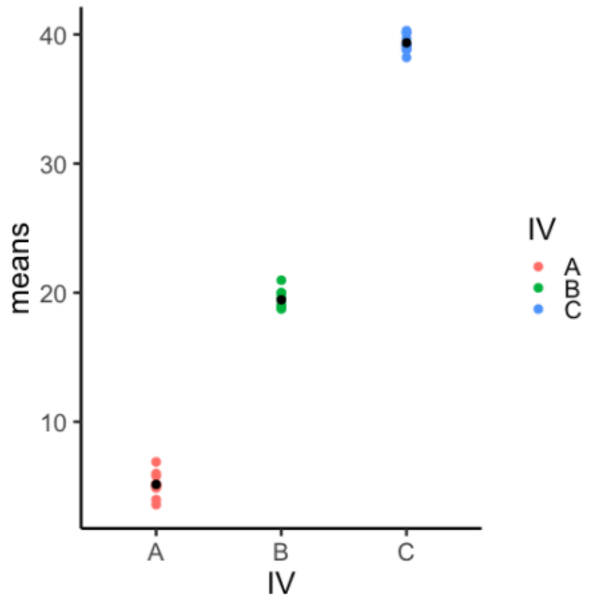
https://csu-statistics.shinyapps.io/ANOVA_oneway_generative/

Try it

- What do you expect F to be when all groups have the same μ's?
  - What do you observe?
  - Now separate them. What happens to F? What happens to p?
  - When is it negative?
- Is the eyeball test generally valid? What do you think? Try?

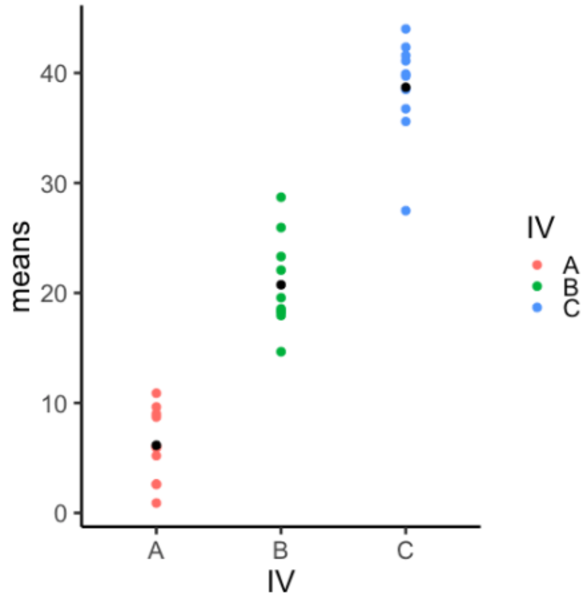
https://csu-statistics.shinyapps.io/ANOVA_oneway_generative/

Example



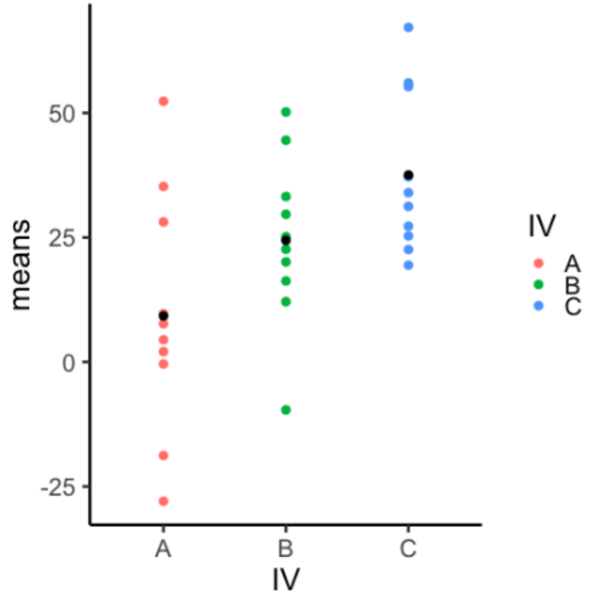
- MS Effect = 2954.08
- MS Error = 0.63
- F = 4720.71

Example



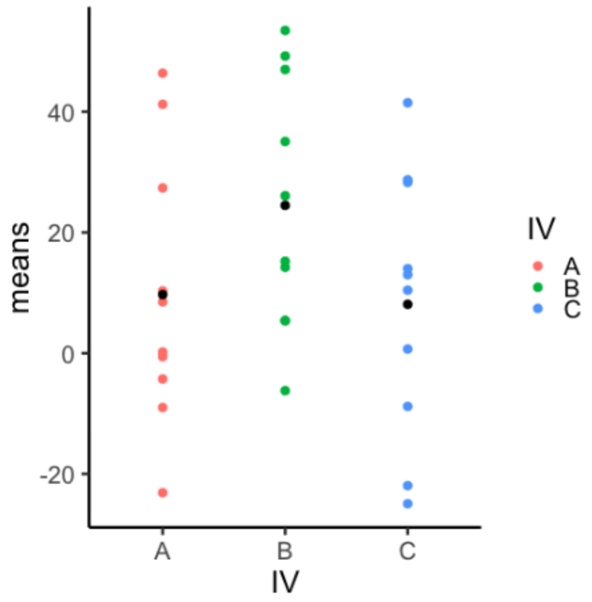
- MS Effect = 2657.61
- MS Error = 17.11
- F = 155.35

Example



- MS Effect = 2006.23
- MS Error = 377.6
- F = 5.31

Example



- MS Effect = 815.97
- MS Error = 472.21
- F = 1.73

One-way ANOVA

You know these:

- k group means – $\mu_1, \mu_2, \ldots, \mu_k$
- k samples
  - k sample mean – $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$
  - Sample size – $n_1, n_2, \ldots, n_k$
  - Within group variation

Question: **Are the group means different?**

- Not: "Which ones differ?”
- Not: "By how much?"
- Just: "Is there a difference *somewhere*?"

Follow-up comparisons

- ANOVA only tests the ***omnibus*** question…
  - Are there any differences anywhere?
- Need to conduct additional tests to compare specific means…
  - Questions like:
- Numerous recommendations for the "right" way to do this
  - Simplest: follow-up t-tests
    - T-test on every group pair!
      - Increases Type-1 error rate

Increasing Type-1 error rate

- **One test:** One decision (binary) with error α of messing up
- **Three tests:** three chances to mess up

  Test 1: Group A vs B → reject?
  Test 2: Group A vs C → reject?
  Test 3: Group B vs C → reject?

  - Each has 5% error.
  - And cross all three?
    - ___ % chance you reject **at least one** incorrectly.

- **A fix**
  - Adjust α for each test to keep overall error at 5%.

Post-hoc corrections

- **Bonferroni:**
  - New $\alpha = 0.05 / 3 = 0.017$
  - Use 0.017 for ALL three tests.

- **Holm**
  - Sequential adjustment for p-values
    - Smallest p-value: compare to $\alpha' = 0.05/3 = 0.017$
    - Second smallest: compare to $\alpha'' = 0.05/2 = 0.025$
    - Largest: compare to $\alpha''' = 0.05/1 = 0.05$
  - Different threshold for each test

One-way ANOVA

- **Test Statistic**

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}$$

- **Global test**
  - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  - Our F-test tells us if we can reject this
  - **If We Reject H_0 ($p < \alpha$):**
    - We can then investigate specific patterns through:
      - a) Planned Contrasts
      - b) post-hoc tests (Bonferroni, etc.)

Example:
Initial ANOVA: H_0: $\mu_1 = \mu_2 = \mu_3$
↓ (Reject H_0)

Specific Questions:
- Is $\mu_1 > \mu_2$? (Pairwise)
- Is $\mu_1 > (\mu_2 + \mu_3)/2$? (Contrast)

One-way ANOVA types

- Between-subjects (independent groups)
- Within-subjects (repeated measures)

Two-way ANOVA

Between group variation / ***variation within groups***

You know these:

- Two factors – A and B
  - A has i levels
  - B has j levels
- that result in i*j cells
  - Cell means
    - $\bar{X}_{11}, \bar{X}_{12}, \ldots, \bar{X}_{ij}$
  - Cell sizes
    - $n_{11}, n_{12}, \ldots, n_{ij}$
  - Within group variation

Question: **Are there main effects? Is there an interaction?**

Test-statistic

- Earlier

$z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

$z = (\bar{X} - \mu) / SE$

$t = (\bar{X} - \mu) / (s/\sqrt{N})$

Now, comparing k means

$$F = \frac{MS_{between}}{MS_{within}}$$

(comparing k means)

“Does Factor A matter? Factor B? Do they interact?"

Two-way ANOVA

- **Use Case:** Comparing means across groups while considering two different categorical factors and their interaction

- **As a Linear Model:**
  - $Y = b_0 + b_1X_1 + b_2X_2 + b_3(X_1 \times X_2) + \varepsilon$
    - where X's are dummy coded

  - NHST
    - Traditional Form:
      - H_0^A: No main effect of factor A
      - H_0^B: No main effect of factor B
      - $H_0^{A\chi B}$: No interaction between A and B
    - lm() equivalent:
      - H_0^A: $b_1 = 0$
      - H_0^B: $b_2 = 0$
      - $H_0^{A\chi B}$: $b_3 = 0$

Two-way ANOVA types

- Between-subjects (both factors between)
- Within-subjects (both factors repeated)
- Mixed (one between, one within)

ANOVA ↔ Experimental Design

ANOVA language → Design language

- Factor → Independent variable
- Level → Condition
- Between-subjects → Independent groups
- Within-subjects → Repeated measures
- One-way → One IV
- Two-way → Two IVs
- Mixed → Mixed design

ANOVA ↔ Experimental Design

ANOVA language → Design language

- Factor → Independent variable
- Level → Condition
- Between-subjects → Independent groups
- Within-subjects → Repeated measures
- One-way → One IV
- Two-way → Two IVs
- Mixed → Mixed design

Often true. But many nuances:

- Observational / quasi-experimental studies
- Selection / Assignment matters in designs
- Random vs fixed effects (same design)
- “nuisance variables” controlled for (blocking)

Three-way ANOVA

You know these:

- Three factors – A, B, C
  - A has i levels
  - B has j levels
  - C has k levels
- that result in i*j*k cells
  - Cell means
    - X $\bar{X}_{111}, \bar{X}_{112}, \ldots, \bar{X}_{ijk}$
  - Cell sizes
    - $n_{111}, n_{112}, \ldots, n_{ijk}$
  - Within cell variation

Question: **Are there main effects? Is there an interaction?**

*Between group variation / **variation within groups***

Test-statistic

- Earlier

$z = (\bar{X} - \mu) / (\sigma/\sqrt{N})$

$z = (\bar{X} - \mu) / SE$

$t = (\bar{X} - \mu) / (s/\sqrt{N})$

Now, comparing k means

$$F = \frac{MS_{between}}{MS_{within}}$$

(comparing k means)

"Does Factor A matter? B? C? Do AB interact? AC? BC? ABC?"

Three-way ANOVA

- Types
  - All between (BBB)
  - All within (WWW)
  - **Mixed designs:**
    - Two between, one within (BBW)
    - One between, two within (BWW)

N-way ANOVA ⇔ Factorial ANOVA

- N factors
- Types
  - Each factor is either between or within.
    - 2^n combinations or types