

# **Unit III**

ECON 3406

**Dr. Josh Martin**

## Differences of Two Proportions

- Like with  $\hat{p}$ , the difference of two sample proportions  $\hat{p}_1 - \hat{p}_2$  can be modeled using a normal distribution (when conditions are met).

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- This standard error comes from the fact that variances of independent variables **add**, even when subtracting.

## Differences of Two Proportions: Standard Errors

- When we talk about the spread of an estimate, we're really talking about **variance** (the square of the standard error).
- If two random variables **A** and **B** are independent, then:

$$\text{Var}(A - B) = \text{Var}(A) + \text{Var}(B)$$

- This might seem counterintuitive – but remember:
  - Even if you're subtracting two noisy measurements, the **uncertainty (noise)** from both still adds up.
  - Think of it like using two shaky rulers. Subtracting doesn't cancel the shakiness – it just combines it!

## Differences of Two Proportions: Simulation Setup

We'll use simulation to understand the sampling distribution of sample proportions for two independent groups.

- Group 1 has a true proportion  $p_1 = 0.5$
- Group 2 has a true proportion  $p_2 = 0.4$
- Each group has  $n = 500$  individuals per sample
- We'll repeat this sampling 1,000 times to observe variation in sample means

## Differences of Two Proportions: Simulation

```
1 set.seed(1)          # ensures reproducibility
2 B <- 1000            # number of simulations
3 n <- 500              # sample size per group
4 p1 <- 0.5             # true proportion in group 1
5 p2 <- 0.4             # true proportion in group 2
6
7 # Create empty vectors to store simulated means and SDs
8 mean_x1 <- mean_x2 <- numeric(B)
9 sd_x1 <- sd_x2 <- numeric(B)
10
11 # Loop to simulate samples for both groups
12 for (i in 1:B) {
13     # Generate random binary outcomes for group 1 (successes)
14     x1 <- rbinom(n, size = 1, prob = p1)
15     mean_x1[i] <- mean(x1)    # sample proportion for group 1
16     sd_x1[i]   <- sd(x1)      # sample SD for group 1
17
18 # Repeat for group 2
19 x2 <- rbinom(n, size = 1, prob = p2)
```

## Comparing Theoretical and Empirical Standard Errors

- We can compare:
  - the **theoretical** standard error (from the formula)
  - the **empirical** standard error (from our simulations)

## Comparing Theoretical and Empirical Standard Errors

```
1 # Theoretical standard error for a sample proportion  
2 sqrt(p1*(1-p1)/n)
```

```
[1] 0.02236068
```

```
1 # Empirical standard error from simulated data  
2 mean(sd_x1) / sqrt(n)
```

```
[1] 0.0223613
```

- The first line gives the theoretical SE:  $\sqrt{p(1 - p)/n}$
- The second line gives the empirical SE, based on simulated SDs
- These values should be nearly identical, validating the normal approximation for large  $n$

## Sampling Distribution for One Group ( $p_1 = 0.5$ )

- We can visualize the distribution of sample proportions across simulations, and overlay a 95% confidence interval around the true mean.

## **Sampling Distribution for One Group ( $p_1 = 0.5$ )**

---

## Sampling Distribution for One Group ( $p_1 = 0.5$ )

- Roughly 5% of simulated sample proportions should fall outside this interval – confirming the 95% confidence level's interpretation.

```
1 # Calculate the proportion of estimates that fall outs
2 mean(ifelse(mean_x1 >= ub_x1 | mean_x1 <= lb_x1, 1, 0))
```

```
[1] 0.052
```

## **Sampling Distribution for One Group ( $p_2 = 0.4$ )**

---

## Sampling Distribution for One Group ( $p_2 = 0.4$ )

- Again, around 5% of simulated estimates will fall outside the interval.
- The spread is slightly narrower than for Group 1 because the variance is smaller.

```
1 # Share of points outside the 95% CI  
2 mean(ifelse(mean_x2 >= ub_x2 | mean_x2 <= lb_x2, 1, 0))
```

```
[1] 0.048
```

## Combining Two Proportions

- Now that we understand the sampling variation of each group separately, we can combine them just as we would when estimating a **difference in proportions**:

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2}$$

- This formula reflects that variances add, even though we're subtracting proportions.

## Simulated Differences Between Groups

```
1 # Compute simulated differences
2 diff <- mean_x1 - mean_x2
3
4 # Theoretical SE for the difference in proportions
5 sqrt(p1*(1-p1)/n + p2*(1-p2)/n)
```

```
[1] 0.03130495
```

```
1 # Empirical SE estimate (not exact but illustrative)
2 mean(sd_x1 + sd_x2) / sqrt(n + n)
```

```
[1] 0.03129505
```

## Simulated Differences Between Groups

---

## Simulated Differences Between Groups

```
1 # Check coverage rates for both formulas  
2 mean(ifelse(diff >= ub_diff_1 | diff <= lb_diff_1, 1,
```

```
[1] 0.037
```

```
1 mean(ifelse(diff >= ub_diff_2 | diff <= lb_diff_2, 1,
```

```
[1] 0.793
```

## Differences of two proportions: Example 1

- Consider an experiment involving patients who underwent cardiopulmonary resuscitation (CPR) following a heart attack and were subsequently admitted to a hospital.
  - Patients were randomly assigned to either a **treatment group** (received a blood thinner) or a **control group** (no blood thinner).
  - The outcome of interest was **survival for at least 24 hours**.

## Differences of two proportions: Example 1

	<b>Survived</b>	<b>Died</b>	<b>Total</b>
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

## Differences of two proportions: Example 1

- Create and interpret a **90% confidence interval** of the difference for the survival rates in the CPR study.

- $p_t - p_c = 0.35 - 0.2 = 0.13$

- $SE = \sqrt{\frac{0.35(1 - 0.35)}{40} + \frac{0.22(1 - 0.22)}{50}} \approx 0.095$

- $0.13 \pm 1.645 \times 0.095 = (-0.027, 0.287)$

## Differences of two proportions: Computing in R

```
1 pt <- 14/40
2 pc <- 11/50
3 nt <- 40
4 nc <- 50
5
6 point_est <- pt - pc
7 se <- sqrt((pt * (1 - pt) / nt) + (pc * (1 - pc) / nc)
8
9 z <- data.frame(
10   sig_level = c(0.01, 0.05, 0.1),
11   z_score = c(2.45, 1.96, 1.645)
12 )
13
14 z$min <- point_est - z$z_score * se
15 z$max <- point_est + z$z_score * se
16 z
```

	sig_level	z_score	min	max
1	0.01	2.450	-0.10396538	0.3639654
2	0.05	1.960	-0.05717230	0.3171723
3	0.10	1.645	-0.02709104	0.2870910

## Differences of two proportions: Visualizing Confidence Intervals

---

## Differences of two proportions: Interpretation

- We are **90% confident** that blood thinners change the 24-hour survival rate by between -3 and 29 percentage points for patients similar to those in the study.
- Because **0% is within this range**, the evidence is inconclusive – we cannot determine whether blood thinners help or harm heart attack patients who have undergone CPR.

## Differences of Two Proportions: Example 2

- A 5-year clinical trial evaluated whether **fish oil supplements** reduce the risk of **heart attacks**.
- Each participant was randomly assigned to one of two groups:
  - **Fish Oil group**
  - **Placebo group**
- We'll examine heart attack outcomes across both groups.

## Differences of Two Proportions: Example 2

<b>Group</b>	<b>Heart Attack</b>	<b>No Event</b>	<b>Total</b>
Fish Oil	145	12,788	12,933
Placebo	200	12,738	12,938

## Differences of Two Proportions: Example 2

- Construct a **95% confidence interval** for the effect of fish oil on heart attack incidence among patients represented by this study.
- Interpret the interval in context:
  - What does the direction and width of the interval suggest?
  - Is there evidence that fish oil has a meaningful effect on heart attack risk?

## Differences of two proportions: Computing in R

```
1 nt <- 12933
2 nc <- 12938
3
4 pt <- 145 / nt
5 pc <- 200 / nc
6
7 point_est <- pt - pc
8 se <- sqrt((pt * (1 - pt) / nt) + (pc * (1 - pc) / nc))
9
10 z <- data.frame(
11   sig_level = c(0.01, 0.05, 0.1),
12   z_score = c(2.45, 1.96, 1.645)
13 )
14
15 z$min <- point_est - z$z_score * se
16 z$max <- point_est + z$z_score * se
```

## Differences of two proportions: Visualizing Confidence Intervals

---

## Differences of two proportions: Interpretation

- The **point estimate** for the effect of fish oil is approximately **-0.0043**, meaning heart attacks occurred **0.43 percentage points less often** in the fish-oil group than in the placebo group.
- We are **90% confident** that fish oil changes the heart-attack rate by between **-0.66 and -0.19 percentage points** for patients similar to those in the study.
- Because this interval **does not include 0**, the reduction in heart-attack risk is **statistically significant** at the 10% (and even 5% and 1%) level.

## Practical vs. Statistical Significance

- While statistically significant, the **effect size is extremely small** – roughly **0.4 fewer heart attacks per 100 individuals**.
- In a large clinical sample, even minor effects can reach significance if variability is low.
- From a **practical** standpoint, such a small reduction may **not justify** the cost, side effects, or adherence burden of treatment.

## More on Two-Proportion Hypothesis Tests

- When conducting a two-proportion hypothesis test, the null hypothesis is typically:  $H_0: p_1 - p_2 = 0$
- However, there are cases where we may want to test for a *specific* difference other than zero.
  - For example, suppose we want to test whether:  $H_0: p_1 - p_2 = 0.10$
- In contexts like these, we use the sample proportions  $\hat{p}_1$  and  $\hat{p}_2$  to check the success-failure condition and to construct the standard error.

## Differences of Two Proportions: Example 3

- A drone quadcopter company is considering a new manufacturer for rotor blades.
- The new manufacturer is more expensive but claims that their higher-quality blades are **3% more reliable**, meaning that 3% more blades pass inspection compared to the current supplier.
- Set up the appropriate hypotheses for this test:
  - $H_0: p_{\text{highQ}} - p_{\text{standard}} = 0.03$
  - $H_A: p_{\text{highQ}} - p_{\text{standard}} \neq 0.03$

## Differences of Two Proportions: Example 3 (Data)

- A quality control engineer collects samples of 1,000 blades from each manufacturer:
  - Current supplier: 899 blades pass inspection
  - Prospective supplier: 958 blades pass inspection
- Using these data, evaluate the hypotheses above at a significance level of 5%.

## Compute the Point Estimate and Standard Error

```
1 p_us <- 958 / 1000
2 p_them <- 899 / 1000
3 point_est <- p_us - p_them
4
5 # Standard error for independent samples
6 se <- sqrt( p_us * (1 - p_us) / 1000 + p_them * (1 -
7
8 p_us; p_them
```

```
[1] 0.958
```

```
[1] 0.899
```

```
1 point_est
```

```
[1] 0.059
```

```
1 se
```

```
[1] 0.01144705
```

## Compute Confidence Intervals for Various Significance Levels

```
1 z <- data.frame(  
2   sig_level = c(0.01, 0.05, 0.10),  
3   z_score = c(2.45, 1.96, 1.645)  
4 )  
5  
6 z$min <- (point_est - z$z_score * se) - 0.03  
7 z$max <- (point_est + z$z_score * se) - 0.03  
8 z
```

	sig_level	z_score	min	max
1	0.01	2.450	0.0009547225	0.05704528
2	0.05	1.960	0.0065637780	0.05143622
3	0.10	1.645	0.0101695994	0.04783040

# Visualizing Confidence Intervals

---

## Compute and Visualize the z-Statistic

```
1 z <- (point_est - 0.03) / se
2
3 set.seed(1)
4 sim <- rnorm(1000, mean = 0.03, sd = se)
5
6 # Probability of observing a value this extreme or larger
7 1 - mean(ifelse(point_est >= sim, 1, 0))
```

```
[1] 0.004
```

```
1 # p-value (right-tailed)
2 1 - pnorm(z)
```

```
[1] 0.005648044
```

## Visualizing the Sampling Distribution

---

### Example 3: Conclusion

- From the standard normal distribution:
  - The right-tail area is approximately 0.004
  - Doubling for a two-tailed test gives  $p = 0.008$
  - Since  $p = 0.008 < 0.05$ , we **reject the null hypothesis**
- We find statistically significant evidence that the higher-quality blades have a pass rate greater than 3% higher than the standard blades – exceeding the company's claims.

## Chi-Squared Distributions: Introduction

- $\chi$  = the greek letter for “chi” (pronounced like “kai”)
- The  $\chi^2$  distribution is a continuous probability distribution that is widely used in statistical inference.
  - Closely related to the standard normal distribution
- If a variable  $Z$  has the standard normal distribution, then  $Z^2$  has the  $\chi^2$  distribution with one **degree of freedom**

## Chi-Squared Distributions: Histograms

---

## Chi-Squared Distributions: Definition

- If  $Z_1, Z_2, \dots, Z_k$  are independent standard normal variables, then...
  - $Z_1^2 + Z_2^2 + \dots + Z_k^2$
  - ...has a  $\chi^2$  distribution with  $k$  **degrees of freedom**.

## Degrees of Freedom: Concept

- A **degree of freedom (df)** represents the number of **independent pieces of information** available to estimate something.
- Whenever we calculate a statistic, we “use up” some information.
  - For example, once we estimate the sample mean, one data point can be perfectly predicted from the others.
  - So for a sample of size  $n$ , only  $(n - 1)$  observations are *free to vary* when computing the sample variance.

## Degrees of Freedom: Intuition

- Average (mean):  $\bar{x} = \frac{1}{n} \sum_i^n x_i = \frac{x_1 + \dots + x_n}{n}$
- if  $i = 4, x_1 = 8, x_2 = 10, x_3 = 12$  and  $\bar{x} = 10\dots$ 
  - ... then  $x_4 = 10$
- So even though we had four data points, only three were free to vary – the fourth is determined by the mean.
- That's why when calculating the sample variance, we divide by  $n$  instead of  $n - 1$ : one degree of freedom has been “used up” in estimating the mean.

## Degrees of Freedom: Motivation

- **Why it matters:**

- Degrees of freedom tell us **how much independent information** our test or estimate is based on.
- They affect the **shape of sampling distributions** (like  $t$ ,  $F$ , and  $\chi^2$ ), which in turn changes the critical values and p-values we use.
- More degrees of freedom → more information → the distribution becomes narrower and more normal-looking.

- **Big idea:**

- Degrees of freedom link **sample size, uncertainty**, and the **reliability of inference** – they remind us that every time we estimate something, we “spend” information.

## Chi-Squared Distributions: Properties

- mean:  $\mu = k$
- variance:  $\sigma^2 = 2k$
- mode occurs at  $\mu - 2$

```
1 set.seed(1)
2 x <- rnorm(1000, mean = 0, sd = 1)
3 x2 <- x^2
4 mean(x2)
```

```
[1] 1.070115
```

```
1 var(x2)
```

```
[1] 2.291541
```

## Chi-Squared Distributions: Multiple Degrees of Freedom

---

## Chi-Squared Distributions: Multiple Degrees of Freedom

```
1 k <- 2
2 for(i in 1:k){
3   set.seed(i)
4   x <- rnorm(1000, mean = 0, sd = 1)
5   x2 <- x^2
6   if(i == 1){
7     z2 <- x2
8   }else{
9     z2 <- z2 + x2
10  }
11 }
12 mean(z2); k
```

```
[1] 2.103046
```

```
[1] 2
```

```
1 var(z2); 2*k
```

```
[1] 4.281053
```

```
[1] 4
```

## Chi-Squared Distributions: Multiple Degrees of Freedom

```
1 k <- 20
2 for(i in 1:k){
3   set.seed(i)
4   x <- rnorm(1000, mean = 0, sd = 1)
5   x2 <- x^2
6   if(i == 1){
7     z2 <- x2
8   }else{
9     z2 <- z2 + x2
10  }
11 }
12 mean(z2); k
```

```
[1] 20.08944
```

```
[1] 20
```

```
1 var(z2); 2*k
```

```
[1] 40.06536
```

```
[1] 40
```

## Chi-Squared Distributions: Multiple Degrees of Freedom

---

## Chi-Squared Tests

- In this section, we develop a method for assessing a **null model when data are binned into categories**.
- This technique is commonly used in two settings:
  - To determine whether a **sample is representative** of a larger population across several groups.
  - To evaluate whether observed data **follow a specified theoretical distribution** (e.g., normal, geometric, Poisson).
- Both scenarios rely on the same statistical tool – the **chi-squared test** – which allows us to examine **all bins simultaneously**, rather than testing one or two categories in isolation.

## Chi-Squared Tests: Jurors Example

- Consider data from a **random sample of 275 jurors** in a small county.
  - Each juror identified their racial group.
  - We would like to test whether this sample is **racially representative** of the eligible population of jurors (i.e., registered voters).

```
1 x <- data.frame(  
2   race      = c("white", "black", "hispanic", "other"  
3   obs_number = c(205, 26, 25, 19),  
4   exp_share  = c(0.72, 0.07, 0.12, 0.09)  
5 )
```

## Chi-Squared Tests: Jurors Example

```
1 n <- sum(x$obs_number)
2 p <- x$exp_share
3 E <- n * p
4 k <- nrow(x)
5
6 n; x$race; E
```

```
[1] 275
```

```
[1] "white"     "black"      "hispanic"   "other"
```

```
[1] 198.00 19.25 33.00 24.75
```

## Chi-Squared Tests: Jurors Example

- The sample proportions of jurors by race do not perfectly match those in the voter population.
- The question is whether these differences are **large enough** to suggest that the jury was **not randomly selected**.
- We can summarize this as a formal hypothesis test:
  - **$H_0$  (Null Hypothesis):** Jurors are randomly selected; any differences reflect natural sampling variation.
  - **$H_a$  (Alternative Hypothesis):** Jurors are not randomly selected; the differences are too large to attribute to chance alone.

## Chi-Squared Tests: Building the Test Statistic

- In previous hypothesis tests, our test statistic took the general form:

$$\frac{\text{Point Estimate} - \text{Null Value}}{\text{Standard Error of Point Estimate}}$$

- This approach involved two key steps:
  1. Measuring the **difference** between the observed value and what we would expect if the null hypothesis were true.
  2. **Standardizing** that difference using the standard error of the point estimate.

## Chi-Squared Test Statistic (Preview)

- For categorical data, the test statistic combines the **squared, standardized deviations** across all categories:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- Where:
  - $O_i$  = observed count in category  $i$
  - $E_i$  = expected count under the null hypothesis
- This statistic follows a **chi-squared distribution** with  $(k - 1)$  degrees of freedom, where  $k$  is the number of categories.

## Chi-Squared Tests: Calculating X^2

```
1 X2_obs <- sum( (x$obs_number - E)^2 / E )
2 X2_obs
```

```
[1] 5.88961
```

## Chi-Squared Tests: Simulations

```
1 set.seed(1)
2 B <- 200000
3 Y <- t(rmultinom(B, size = n, prob = p))
4 m <- matrix(E, B, k, byrow=TRUE)
5 colnames(m) <- colnames(Y) <- x$race
6
7 head(Y)
```

	white	black	hispanic	other
[1, ]	196	20	29	30
[2, ]	205	23	25	22
[3, ]	203	15	36	21
[4, ]	199	25	32	19
[5, ]	202	16	34	23
[6, ]	203	22	30	20

```
1 head(m)
```

	white	black	hispanic	other
[1, ]	198	19.25	33	24.75
[2, ]	198	19.25	33	24.75
[3, ]	198	19.25	33	24.75
[4, ]	198	19.25	33	24.75
[5, ]	198	19.25	33	24.75
[6, ]	198	19.25	33	24.75

## Chi-Squared Tests: Simulations

- The p-value of 0.117 indicates that, if jurors were selected randomly according to the county's racial composition, there is about an 11.7% chance of observing differences in juror counts at least as extreme as those in the sample

```
1 X2_sim <- rowSums( (Y - m)^2 / m )
2 p_value1 <- mean(X2_sim >= X2_obs)
3 p_value1
```

```
[1] 0.11668
```

## Chi-Squared Tests: Visualizing the p-value

---

## Chi-Squared Tests: Attempting a More Familiar Approach (Incorrect)

```
1 z <- as.data.frame(Y)
2 for(i in 1:4){
3   z[,i] <- (z[,i] - mean(z[,i])) / sd(z[,i])
4 }
5 x2 <- z[,1]^2 + z[,2]^2 + z[,3]^2
6 p_value2 <- mean(ifelse(X2_obs < x2, 1, 0))
7 p_value2
```

[1] 0.13081

## Chi-Squared Tests: Attempting a More Familiar Approach (Incorrect)

---

## Why the "More Familiar Way" Is Incorrect

- This approach standardizes each category as if they were **independent samples**.
  - When one category increases, another must decrease, introducing **negative covariance**.
  - Ignoring this dependence effectively uses **too many degrees of freedom**, overstating variability.
- The correct chi-squared test adjusts for this by using  $(k - 1)$  degrees of freedom.

## Chi-Squared Tests: Stock Market Example

- The chi-squared framework can also be used to **evaluate how well a statistical model fits observed data**.
- Suppose we analyze **daily stock returns from the S&P 500** over a 10-year period to test whether market movements are **independent across days**.
  - In other words, does knowing whether the market went up or down yesterday help predict what happens today?
- This question has clear implications for traders:
  - If past information helps forecast future returns, that knowledge could provide a **trading advantage**.

## Chi-Squared Tests: Stock Market Example

- To explore this, we can label each trading day as “Up” or “Down”, and record the **number of days between Up days** to study the waiting-time pattern.
- If daily market movements are truly **independent**, the waiting time between positive (Up) days should follow a **geometric distribution**.
- The geometric distribution gives the probability of waiting  $k$  trials to observe the first success:  $P(X = k) = (1 - p)^{k-1}p$ 
  - $p$  = probability that a given day is an Up day
  - $X$  = number of days until the next Up day

## Chi-Squared Tests: Computing Expected Counts

```
1 x <- data.frame(
2   days = 1:7,
3   observed = c(717, 369, 155, 69, 28, 14, 10)
4 )
5
6 p <- 0.545
7 x$expected <- round((1-p)^(x$days)-1)*p*sum(x$observed)
8
9 x
```

	days	observed	expected
1	1	717	742
2	2	369	338
3	3	155	154
4	4	69	70
5	5	28	32
6	6	14	14
7	7	10	7

```
1 x2 <- sum( (x$observed - x$expected)^2 / x$expected )
2 x2
```

[1] 5.492007

## Chi-Squared Tests: Comparing Observed and Expected Rallies

---

## Chi-Squared Tests: Simulating the Chi-Squared Distribution

```
1 for(i in 1:6){  
2   set.seed(i)  
3   z <- rnorm(sum(x$observed), mean = 0, sd = 1)  
4   if(i == 1){  
5     y <- z^2  
6   }else{  
7     y <- y + z^2  
8   }  
9 }  
10  
11 mean(ifelse(x2 > y, 1, 0))
```

```
[1] 0.5198238
```

```
1 pchisq(x2, df = 6)
```

```
[1] 0.5175767
```

# Chi-Squared Tests: Visualizing the p-Value

---

## Intro to t-distributions

- We rarely know the true population mean ( $\mu$ ) — we estimate it from a sample.
- When the population standard deviation ( $\sigma$ ) is unknown (which is most of the time), we rely on the  $t$ -distribution instead of the normal model.
- The  $t$ -test helps us quantify uncertainty and make inferences about population parameters using sample data.

## t-distributions

---

## t-distributions: Degrees of Freedom

---

## Intuition of t-distributions

```
1 B <- 10000
2
3 set.seed(1)
4 x <- rnorm(B, mean = 0, sd = 1)
5 dx <- density(x)
6
7 df <- 3
8 set.seed(1)
9 t_vals <- replicate(B, {
10   x <- rnorm(df+1)                      # sample size
11   mean(x) / (sd(x) / sqrt(df+1))        # t-statistic
12 })
13 dt <- density(t_vals)
```

## Using the t-Distribution

- We use a **t-distribution with ( $df = n - 1$ )** to model the sample mean when the sample size is  $n$ .
  - As the number of observations increases, the **degrees of freedom (df)** also increase.
  - With larger df, the **t-distribution more closely resembles the standard normal distribution**.
- When  $df \approx 30$  or greater, the t-distribution is **nearly indistinguishable** from the normal model.
  - In general, the **degrees of freedom describe the shape** of the t-distribution — larger df means the distribution becomes narrower and more normal-like.

## Example: Mercury in Dolphins

- Let's apply the t-distribution in a real-world example involving **mercury content in dolphin muscle**.
  - Elevated mercury concentrations are a serious concern for **both dolphins and humans** who consume them.
  - We'll construct a **confidence interval** for the **average mercury concentration** using a **sample of 19 Risso's dolphins** from the **Taiji area in Japan**.

## Example: Mercury in Dolphins

### Statistic    Value

n	19
$\bar{x}$ (mean)	4.4
s (SD)	2.3
Min	1.7
Max	9.2

$$\text{point estimate} \pm t_{df}^* \times \frac{s}{\sqrt{n}}$$

## t-statistics and p-values

---

## t-statistics and confidence intervals

```
1 B <- 100000
2 n = 19
3 point_est <- 4.4
4 s <- 2.3
5
6 se <- x_int * s /sqrt(n)
7
8 upper_est <- point_est + se
9 lower_est <- point_est - se
10
11 round(c(lower_est, upper_est), 2)
```

```
[1] 3.3 5.5
```

## Example 2: Are Runners Getting Faster?

Is the **typical U.S. runner** getting faster or slower over time?

We'll explore this question using data from the **Cherry Blossom 10-Mile Run**, held each spring in **Washington, D.C.**

- In **2006**, the average finish time for all runners was **93.29 minutes** (about **93 minutes and 17 seconds**).
- We'll use a **sample of 100 participants** from the **2017 race** to test whether the **average finish time has changed**.
- In other words, are runners **getting faster or slower**, or has there been **no meaningful change**?

## Example 2: Calculating the t-statistic

$$t = \frac{\text{estimate} - \text{null}}{sd / \sqrt{n}}$$

```
1 y <- data.frame(
2   mean = 97.32,
3   sd = 16.98,
4   n = 100
5 )
6 df <- y$n - 1
7
8 t_stat <- (y$mean - 93.29) / (y$sd / sqrt(y$n))
9 t_stat
```

[1] 2.37338

## Example 2: Vizualizing the p-value

---

## Example 2: Interpreting the Results

```
1 1 - mean(ifelse(t_stat > x, 1, 0))
```

```
[1] 0.01
```

```
1 round(1 - pt(t_stat, df = df), 3)
```

```
[1] 0.01
```

- Because the **p-value is smaller than 0.05**, we **reject the null hypothesis**.
  - The data provide **strong evidence** that the **average run time in 2017** differs from the **2006 average**.
  - Since the **observed mean time is higher** than the 2006 mean, and we've rejected the null, we conclude that: **Runners were slower on average in 2017 than in 2006.**

## Difference of Two Means

- In this section, we consider the difference between **two population means**  $\mu_1 - \mu_2$  when the data are **not paired**.
  - As with a single sample, we must check certain **conditions** to ensure we can use the **t-distribution**.
  - The **point estimate** for the difference is  $\bar{x}_1 - \bar{x}_2$ .
  - A new **standard error formula** is used to reflect variation across both samples.
  - Aside from these two differences, the procedures are **nearly identical** to the one-mean case.

## Example 3: ESC and Heart Recovery

- Does **treatment using embryonic stem cells (ESCs)** help improve heart function after a heart attack?
  - An experiment tested this question in **sheep** that had experienced a **heart attack**.
  - Each sheep was **randomly assigned** to either an **ESC treatment group** or a **control group**.
  - Researchers measured the **change in heart pumping capacity** after treatment.
  - A **positive value** indicates **stronger recovery**.
  - Our goal is to construct a **95% confidence interval** for the **effect of ESC treatment** on heart recovery compared to the control group.

## **Example 3: Distribution**

---

## Example 3: Summary Statistics

```
1 y <- read.csv("stem_cell.csv")
2 y$diff <- y$after - y$before
3 head(y)
```

```
trmt before after diff
1 ctrl 35.25 29.50 -5.75
2 ctrl 36.50 29.50 -7.00
3 ctrl 39.75 36.25 -3.50
4 ctrl 39.75 38.00 -1.75
5 ctrl 41.75 37.50 -4.25
6 ctrl 45.00 42.75 -2.25
```

```
1 x <- aggregate(diff ~ trmt, y,
2                         function(x) c(mean = mean(x),
3                                         sd = sd(x),
4                                         n = length(x)))
5 x <- as.data.frame(do.call(cbind, x))
6 x[,2:4] <- apply(x[,2:4], 2, as.numeric)
7 x
```

```
trmt      mean       sd n
1 ctrl -4.333333 2.764168 9
2 esc  3.500000 5.172040 9
```

## Example 3: Standard Errors

$$SE = \sqrt{SE_1 + SE_2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

```
1 mean_diff <- x$mean[x$trmt == "esc"] - x$mean[x$trmt =
2 se_c <- x$sd[x$trmt == "ctrl"]^2 / x$n[x$trmt == "ctr"]
3 se_t <- x$sd[x$trmt == "esc"]^2 / x$n[x$trmt == "esc"]
4 se <- sqrt(se_c + se_t)
5 se
```

```
[1] 1.954784
```

```
1 t_stat <- mean_diff / se
2 t_stat
```

```
[1] 4.007263
```

```
1 1 - pt(t_stat, df = 8)
```

```
[1] 0.001954976
```

## Example 4: Smoking and Birth Weight

The `ncbirths` dataset contains a **random sample of 150 mothers and their newborns** in North Carolina collected over the course of a year.

- We focus on two key variables:
  - **weight** – the newborn's birth weight (in pounds or ounces)
  - **smoke** – whether the mother **smoked during pregnancy**
- Our question: Is there **convincing evidence** that newborns of mothers who smoke have a **different average birth weight** than those of non-smoking mothers?

## Example 4: Randomization

```
1 y <- read.csv("ncbirths.csv")
2
3 set.seed(1)
4 y$random_number <- rnorm(nrow(y), mean = 0, sd = 1)
5 y <- y[order(y$random_number),]
6
7 t <- y[y$habit == "smoker",]
8 c <- y[y$habit == "nonsmoker",]
9
10 y <- rbind(t[1:50,],
11             c[1:100,])
```

## Example 4: Summary Statistics

```
1 x <- aggregate(weight ~ habit, y,
2                         function(x) c(mean = mean(x),
3                                         sd = sd(x),
4                                         n = length(x)))
5 x <- as.data.frame(do.call(cbind, x))
6 x[,2:4] <- apply(x[,2:4], 2, as.numeric)
7 x
```

	habit	mean	sd	n
1	nonsmoker	7.4757	1.446375	100
2	smoker	6.9136	1.461228	50

## Example 4: t-test and p-value

```
1 mean_diff <- x$mean[x$habit == "smoker"] - x$mean[x$habit == "nonsmoker"]
2 mean_diff
```

```
[1] -0.5621
```

```
1 se_c <- x$sd[x$habit == "nonsmoker"]^2 / x$n[x$habit == "nonsmoker"]
2 se_t <- x$sd[x$habit == "smoker"]^2 / x$n[x$habit == "smoker"]
3 se <- sqrt(se_c + se_t)
4 se
```

```
[1] 0.2522375
```

```
1 t_stat <- mean_diff / se
2 t_stat
```

```
[1] -2.228456
```

```
1 pt(t_stat, df = min(x$n))
```

```
[1] 0.01518814
```

## **ANOVA: Comparing More Than Two Means**

- Sometimes we want to compare **means across multiple groups**.
- We might be tempted to make **pairwise comparisons** – for example, comparing
  1. Group 1 vs. Group 2
  2. Group 1 vs. Group 3
  3. Group 2 vs. Group 3

## **ANOVA: Comparing More Than Two Means**

- However, this approach can be **misleading**. With many groups, making multiple comparisons **increases the chance of finding a difference by random chance**, even if no real differences exist.
  - Instead, we use a **holistic test** – the **Analysis of Variance (ANOVA)** – to assess whether there is **evidence that at least one group mean differs from the others**.
- In short: **ANOVA saves the day** by allowing us to test for **overall mean differences** across several groups at once.

## Core Idea of ANOVA

- We will learn a new method called **analysis of variance** (ANOVA) and a new test statistic called  $F$
- ANOVA uses a single hypothesis test to check whether the means across many groups are equal
  - $H_0 = \mu_1 = \mu_2 = \dots = \mu_k$
  - $H_A = \text{at least one mean is different}$

## ANOVA Example: Exam Scores Across Classes

- College departments often offer **multiple sections of the same introductory course** each semester to meet high student demand.
  - Suppose a **statistics department** runs **three lectures** (A, B, and C) of an introductory statistics course.
- We want to know whether there are **statistically significant differences** in the **average first exam scores** across these three classes.
  - ANOVA allows us to test whether **any of the class means differ**, rather than relying on multiple pairwise comparisons.

## ANOVA: SSG

Statisticians measure variation using squared differences. The **sum of squares for the groups** (SSG) measures the differences between the group means  $\bar{x}_i$  and the grand mean  $\bar{x}$  like so:

$$SSG = \sum_{i=1}^k n_i(\bar{x}_i - \bar{x})^2$$

## ANOVA: SSG

```
1 a <- c(90, 87, 88, 78)
2 b <- c(87, 85, 80)
3 c <- c(95, 93, 90, 88, 85)
4 na <- length(a)
5 nb <- length(b)
6 nc <- length(c)
7 ua <- mean(a)
8 ub <- mean(b)
9 uc <- mean(c)
10 u <- mean(c(a, b, c))
11 SSG <- na*(ua - u)^2 + nb*(ub - u)^2 + nc*(uc - u)^2
12 SSG
```

[1] 84.11667

## ANOVA: MSG

The **mean square between groups** (MSG) is calculated using both the groups' degrees of freedom  $df_G$  and the sum of squares between groups SSG like so:

$$MSG = \frac{1}{df_G} SSG$$

```
1 k <- 3
2 df1 <- k - 1
3
4 MSG <- SSG / df1
5 MSG
```

```
[1] 42.05833
```

## ANOVA: SSE

- The **sum of squared errors** (SSE) is the sum of the squares of the differences between each sample's observations with each respective sample mean
- The calculation is simplified by first finding the variance  $s_i^2$  of each sample:

```
1 a <- c(90, 87, 88, 78)
2 b <- c(87, 85, 80)
3 c <- c(95, 93, 90, 88, 85)
4 sa <- sd(a)
5 sb <- sd(b)
6 sc <- sd(c)
7 sa; sb; sc
```

```
[1] 5.315073
```

```
[1] 3.605551
```

```
[1] 3.962323
```

## ANOVA: SSE

$$SSE = \sum_{i=1}^k (n_i - 1)s_i^2$$

```
1 a <- c(90, 87, 88, 78)
2 b <- c(87, 85, 80)
3 c <- c(95, 93, 90, 88, 85)
4 sa <- sd(a)
5 sb <- sd(b)
6 sc <- sd(c)
7 na <- length(a)
8 nb <- length(b)
9 nc <- length(c)
10 SSE <- (na - 1)*sa^2 + (nb - 1)*sb^2 + (nc - 1)*sc^2
11 SSE
```

[1] 173.55

## ANOVA: MSE

$$MSE = \frac{1}{df_E} SSE$$

```
1 n <- length(c(a, b, c))
2 k <- 3
3 MSE <- SSE / (n - k)
4 MSE
```

```
[1] 19.28333
```

## ANOVA: F-Statistic

$$F = \frac{MSG}{MSE}$$

- We now have comparable measures of variability both between the groups (MSG) and within the groups (MSE).
  - If variation among group means is simply random, F is typically close to 1

```
1 f_stat <- MSG / MSE  
2 f_stat
```

```
[1] 2.181072
```

## ANOVA: p-value from the F-statistic

```
1 set.seed(1)
2 x <- rf(1e6, df1 = k - 1, df2 = n - k)
3 dx <- density(x)
4 1 - mean(ifelse(f_stat > x, 1, 0))
```

```
[1] 0.169123
```

## ANOVA: Vizualizing the F-statistic

---

## ANOVA: p-value from the F-statistic (easy)

```
1 z <- data.frame(  
2   score = c(a, b, c),  
3   group = c(rep('a', na), rep('b', nb), rep('c', nc))  
4 )  
5 z
```

	score	group
1	90	a
2	87	a
3	88	a
4	78	a
5	87	b
6	85	b
7	80	b
8	95	c
9	93	c
10	90	c
11	88	c
12	85	c

```
1 summary(aov(score ~ group, z))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	84.12	42.06	2.181	0.169
Residuals	9	173.55	19.28		

## ANOVA: Statistical Significance

```
1 y <- read.csv("exam_grades.csv")
2 head(y)
```

	semester	sex	exam1	exam2	exam3	course_grade
1	2000-1	Man	84.5	69.5	86.5	76.2564
2	2000-1	Man	80.0	74.0	67.0	75.3882
3	2000-1	Man	56.0	70.0	71.5	67.0564
4	2000-1	Man	64.0	61.0	67.5	63.4538
5	2000-1	Man	90.5	72.5	75.0	72.3949
6	2000-1	Man	74.0	78.5	84.5	71.4128

```
1 summary(aov(course_grade ~ semester, y))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
semester	5	1261	252.14	2.719	0.0208 *						
Residuals	227	21053	92.74								
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

