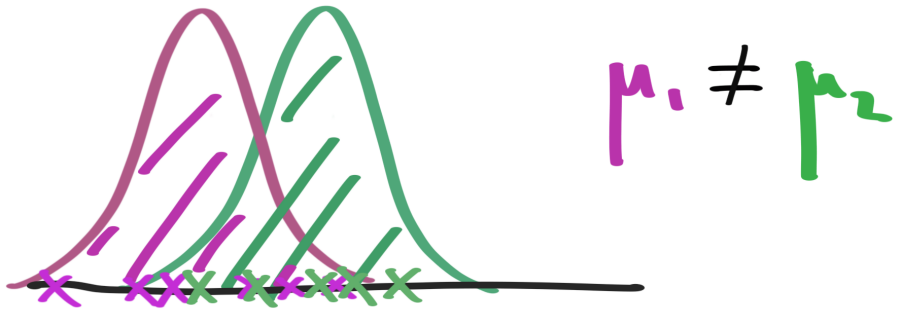
Hypothesis Testing



Statistical Experimental Design

Kris Sankaran | UW Madison | 16 September 2021

Motivation

- Experimental design is concerned with the way many factors (each with many levels) can affect a response
- Simplification: One factor with only two levels
- This still requires key conceptual developments,
  - Hypothesis testing
  - p-values
  - Confidence intervals
  - Power analysis
  - Model-checking

Settings

- Two-sample testing is also important in its own right
- The two sample t-test is probably the most widely used statistical procedure from the last 100 years
- Famous applications: Clinical trials, Internet A/B Tests

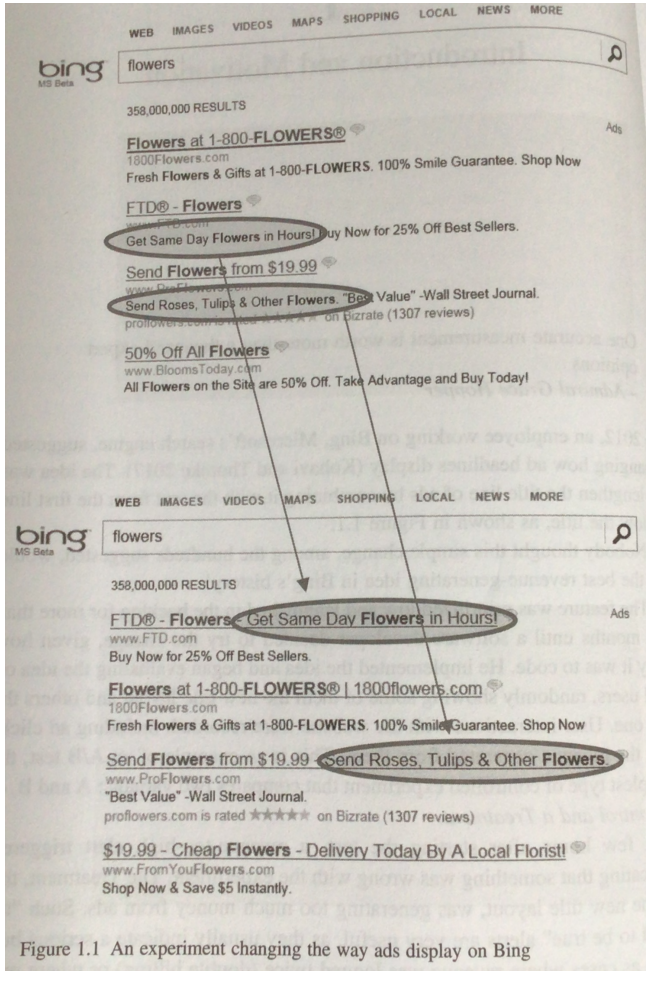
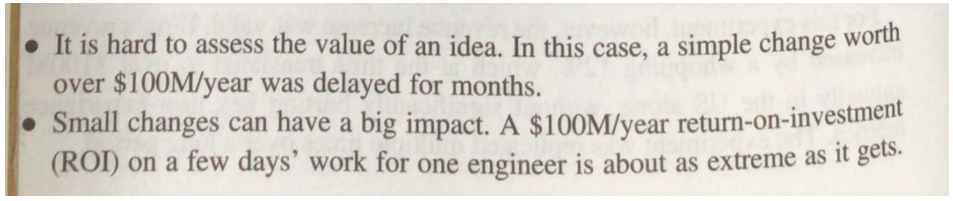


Figure 1.1 An experiment changing the way ads display on Bing

Settings

- Two-sample testing is also important in its own right
- The two sample t-test is probably the most widely used statistical procedure from the last 100 years
- Famous applications: Clinical trials, Internet A/B Tests



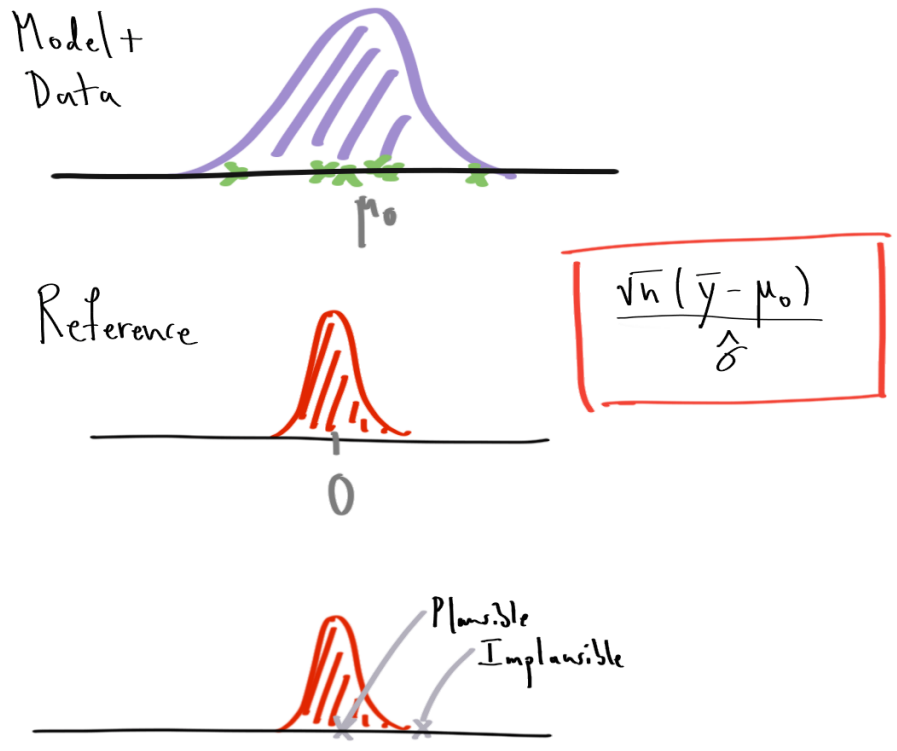
- It is hard to assess the value of an idea. In this case, a simple change worth over $100M/year was delayed for months.
- Small changes can have a big impact. A $100M/year return-on-investment (ROI) on a few days' work for one engineer is about as extreme as it gets.

Philosophy

- Goal: How can we make general conclusions, based only on specific evidence?
- Main idea:
  - Propose a model of the world (the null hypothesis).
  - See whether data you collect is consistent with that theory

Recipe

- Pose a null hypothesis about the world
- Define a test statistic that should detect departures from that null hypothesis
- Determine that statistic's reference distribution
- Compute the test statistic on real data, and see if it's plausibly from the reference



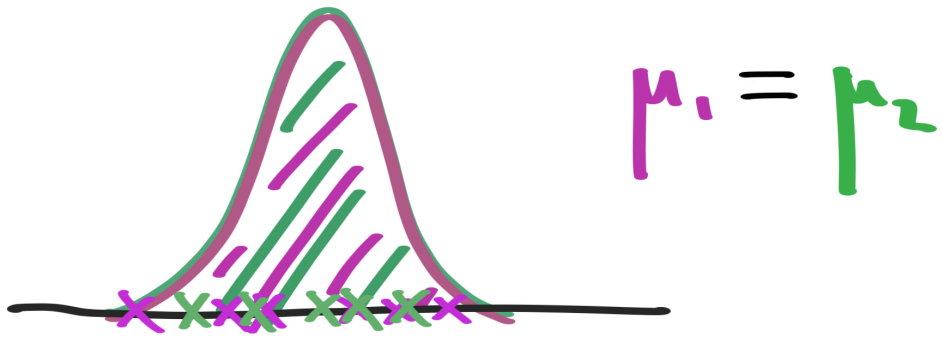
Example: Concrete Mortars

- Compare two mortars
- We assume

$$y_1^{(1)}, \ldots, y_{n_1}^{(1)} \sim \mathcal{N}(\mu_1, 1)$$
$$y_1^{(2)}, \ldots, y_{n_2}^{(2)} \sim \mathcal{N}(\mu_2, 1)$$

  for some unknown μ_1 and μ_2.
- The null and alternative hypotheses are,

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$



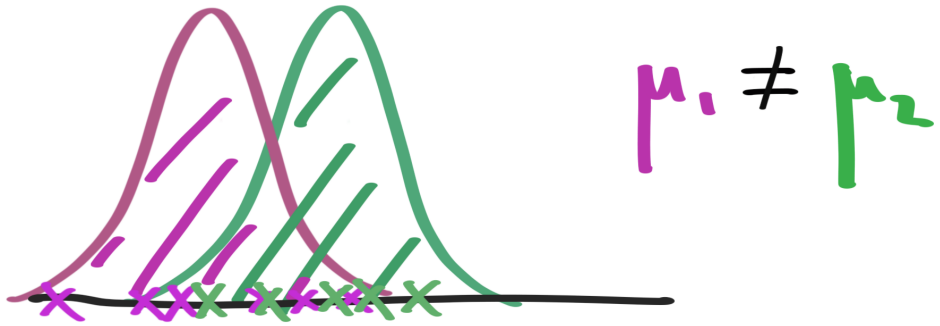
Example: Concrete Mortars

- Compare two mortars
- We assume

$$y_1^{(1)}, \ldots, y_{n_1}^{(1)} \sim \mathcal{N}(\mu_1, 1)$$
$$y_1^{(2)}, \ldots, y_{n_2}^{(2)} \sim \mathcal{N}(\mu_2, 1)$$

  for some unknown μ_1 and μ_2.
- The null and alternative hypotheses are,

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$



Test Statistic

- We are free to choose a test statistic. Let's consider,

$$T\left(\mathbf{y}^1, \mathbf{y}^2\right) = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{2}{5}}}$$

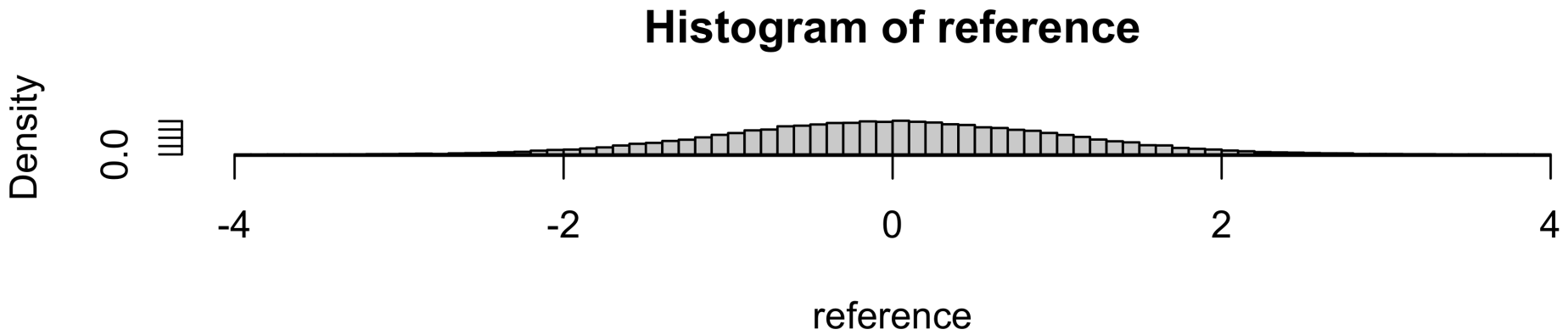
- The statistic is standardized. Under the null, the statistic has mean 0 and variance 1. This follows from,

$$\begin{aligned} \operatorname{Var}\left(\bar{y}_1 - \bar{y}_2\right) &= \operatorname{Var}\left(\bar{y}_1\right) + \operatorname{Var}\left(\bar{y}_2\right) \\ &= \frac{1}{5} + \frac{1}{5} \end{aligned}$$

Simulating a Reference Distribution

```
test_stat <- function(y1, y2, n = 5) {
  (mean(y1) - mean(y2)) / sqrt(2 / n)
}

reference <- rerun(5e4, test_stat(rnorm(5), rnorm(5))) %>%
  unlist()
hist(reference, breaks = 100, freq = FALSE)
```



Reference Distribution

We could have found the reference distribution theoretically, because it is a linear combination of (independent) Gaussians.

```
hist(reference, breaks = 100, freq = FALSE)
lines(seq(-5, 5, .1), dnorm(seq(-5, 5, .1)), col = "red", lw = 1)
```

Plausibility on Real Data

In the final step, we compute the test statistic on real data and see whether the value is plausible, with respect to the reference.

```
real_data <- data.frame(
  mortar = c(rep("A", 5), rep("B", 5)),
  y = c(rnorm(5), rnorm(5, 1.5))
)

ggplot(real_data) +
  geom_point(aes(mortar, y))
```

Plausibility on Real Data

- The magenta is the observed test statistic. What is blue?

```
t_observed <- test_stat(real_data$y[1:5], real_data$y[6:10])
hist(reference, breaks = 100, freq = FALSE)
lines(seq(-5, 5, .1), dnorm(seq(-5, 5, .1)), col = "red", lw = 1)
abline(v = t_observed, col = "magenta", lw = 3)
abline(v = qnorm(0.975), col = "blue", lw = 3)
abline(v = qnorm(0.025), col = "blue", lw = 3)
```

Plausibility on Real Data

- p-values are a measure of the plausibility of the observed statistic.
- It is the probability of observing (under the reference) a test-statistic as or more extreme than the one we did observe

```
2 * (pnorm(t_observed))
```

```
## [1] 0.0003080053
```

Package Implementation

This type of two-sample test where the variances are known is called a z-test.

```
library(BSDA)
library(broom)
z.test(real_data$y[1:5], real_data$y[6:10], "two.sided", sigma.x = 1, sigma.y = 1) %>%
  tidy()
```

```
## # A tibble: 1 × 8
##   estimate1 estimate2 statistic  p.value conf.low conf.high
##       <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1    -0.893      1.39     -3.61 0.000308    -3.52     -1.04
## # … with 2 more variables: method <chr>, alternative <chr>
```

t Tests

Estimating σ^2

What if we hadn't known that variance of the y_i?

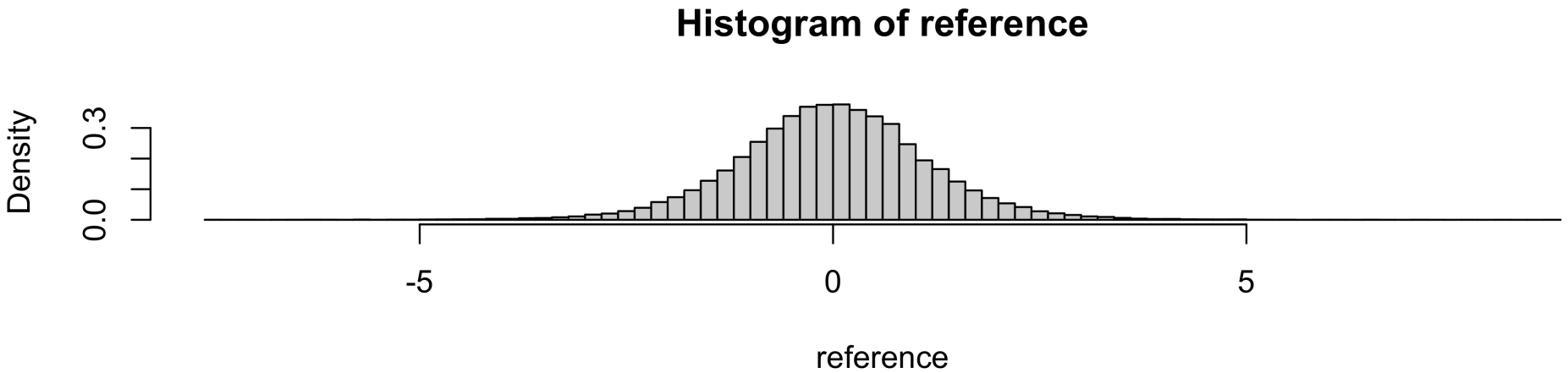
We could try plugging-in an estimate,

$$T\left(\mathbf{y}^1, \mathbf{y}^2\right) = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{2\hat{\sigma}^2}{5}}}$$

Simulating Reference Distribution

```
test_stat <- function(y1, y2, n = 5) {
  sigma <- sqrt((var(y1) + var(y2)) / 2)
  (mean(y1) - mean(y2)) / (sigma * sqrt(2 / n))
}

reference <- rerun(5e4, test_stat(rnorm(5), rnorm(5))) %>%
  unlist()
hist(reference, breaks = 100, freq = FALSE)
```



Exercise

These distributions don't quite match...

- Does this fact have any consequences for testing?
- Why is this reference distribution different anyways?

```
hist(reference, breaks = 100, freq = FALSE)
lines(seq(-5, 5, .1), dnorm(seq(-5, 5, .1)), col = "red", lw = 1)
```

t-test

- Since the true reference's tails are heavier than Normal, our p-values would be over-optimistic
- The issue is that $\hat{\sigma}$ is itself a random quantity

```
hist(reference, breaks = 100, freq = FALSE)
lines(seq(-5, 5, .1), dt(seq(-5, 5, .1), df=8), col = "red", lw = 1)
```

Direct Implementation

- We can use the t-distribution as a reference, and then compute test results and p-values as before.
- This is implemented by the `t.test` function in R

```
t.test(real_data$y[1:5], real_data$y[6:10], var.equal = TRUE) %>%
  tidy()
```

```
## # A tibble: 1 × 10
##   estimate estimate1 estimate2 statistic p.value parameter
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>
## 1    -2.28    -0.893      1.39     -4.21 0.00294         8
## # … with 4 more variables: conf.low <dbl>, conf.high <dbl>,
## #   method <chr>, alternative <chr>
```

Confidence Intervals

Definition

- Hypothesis testing seems a bit roundabout.
- A 95% confidence interval is a (random) interval $[L, U]$ satisfying

$$\mathbf{P}\left(\theta \in [L, U]\right) = 0.95$$

Computation

- For the two-sample testing problem, our earlier code returned intervals for the difference $\bar{y}_1 - \bar{y}_2$
- If you know the distribution for a test statistic, you can often back out an associated confidence interval

```
t.test(real_data$y[1:5], real_data$y[6:10], var.equal = TRUE) %>%
  tidy() %>%
  select(starts_with("conf"))
```

```
## # A tibble: 1 × 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1    -3.53     -1.03
```

Example

The true difference between the groups is 0.

```
intervals <- rerun(500, tidy(t.test(rnorm(10), rnorm(10)))) %>%
  bind_rows(.id = "id") %>%
  mutate(false_alarm = (conf.low > 0) | (conf.high < 0)) %>%
  select(id, starts_with("conf"), false_alarm)
head(intervals, 3)
```

```
## # A tibble: 3 × 4
##   id    conf.low conf.high false_alarm
##   <chr>    <dbl>     <dbl> <lgl>
## 1 1       -0.923     0.782 FALSE
## 2 2       -0.814     1.42  FALSE
## 3 3       -1.17      0.725 FALSE
```

```
mean(intervals$false_alarm)
```

```
## [1] 0.048
```

Example

```
ggplot(intervals) +
  geom_segment(aes(id, conf.low, col = false_alarm, xend = id, yend = conf.high)) +
  scale_color_brewer(palette = "Set2") +
  labs(y = "Confidence Interval") +
  theme(axis.text.x = element_blank(), panel.grid.major.x = element_blank())
```

Derivation (if time)

Define the quantities,

$$
\begin{aligned}
T\left(\mathbf{y}^1, \mathbf{y}^2\right) &:= \bar{y}_1 - \bar{y}_2 \\
\theta &:= \mu_1 - \mu_2 \\
\hat{\sigma} &:= S_p \sqrt{\frac{2}{n}} \\
t_{0.025,2(n-1)} &:= t_{\text{left}} \\
t_{0.975,2(n-1)} &:= t_{\text{right}}
\end{aligned}
$$

Derivation (if time)

The centered and scaled difference in means is t-distributed,

$$\mathbf{P}\left(\frac{T(y)-\theta}{\hat{\sigma}} \in \left[t_{\text{left}}, t_{\text{right}}\right]\right) = 0.95$$

so rearranging terms and using the fact $t_{\text{left}} = -t_{\text{right}}$,

$$\mathbf{P}\left(\theta \in \left[T(y) - \hat{\sigma}t_{\text{right}}, T(y) + \hat{\sigma}t_{\text{right}}\right]\right) = 0.95$$

i.e., this interval is a confidence interval.

Exercise

Compute a 95% confidence interval for the difference in means of samples from group A vs. B.

```
obs <- data.frame(
  group = c(rep("A", 8), rep("B", 8)),
  y = c(0.97, 1.28, -0.29, -0.20, -1.43, 0.67, -0.14, 0.81, 3.51, 1.55, 1.38, 1.54, 3.17, 2.
)
```

Diagnostics and Power

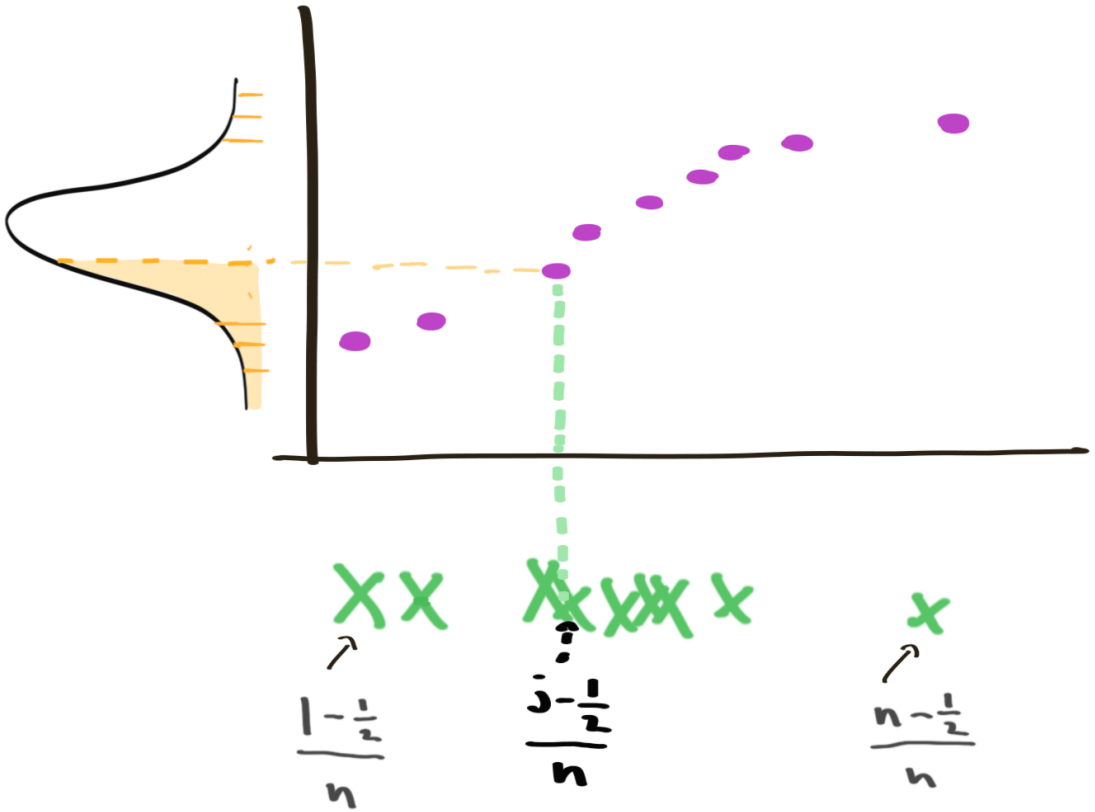
Assumptions

Our tests depend on three assumptions,

- Samples are independent
- The standard deviations are equal
- The populations are normally distributed

QQ Plots

We can check the second two assumptions using Quantile-Quantile (a.k.a. normal probability) plots



Actually Normal

```
qqnorm(rnorm(1000))
abline(0, 1, col = "red")
```

Heavy Tails

```
qqnorm(rt(1000, 2))
abline(0, 1, col = "red")
```

Exercise

Without making a histogram, determine this distribution's shape.

```
qqnorm(scale(exp(rnorm(100))))
abline(0, 1, col = "red")
```

Power

- The power of a test is the probability it can detect a difference when one exists
- It depends on a few things: sample size, effect size, and test statistic used
- Sometimes you can compute this probability by hand, but often it will be necessary to simulate