PLSC 502 – Fall 2023
Two-Group Comparisons
(+ Power)

October 30, 2023

Association, Part I

Association = two variables have nonzero covariance...

- Valuable for **description**
- Starting point for **explanation**
    - "Correlation is not causation, but it sure is a hint." (attributed to E. Tufte)
    - Particularly valuable in experiments / quasi-experiments
    - Is also often a *starting point* for thinking about **model specification**
- Obviously also important for **prediction**

“Student’s” t...

“...the T-Distribution, also known as Student’s t-distribution, gets its name from William Sealy Gosset who first published it in English in 1908 in the scientific journal *Biometrika* using the pseudonym “Student” because his employer preferred staff to use pen names when publishing scientific papers. Gosset worked at the Guinness Brewery in Dublin, Ireland, and was interested in the problems of small samples – for example, the chemical properties of barley with small sample sizes.

Gosset had been hired due to Claude Guinness’s policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness’s industrial processes. Gosset devised the t-test as an economical way to monitor the quality of stout. The Student’s t-test work was submitted to and accepted in the journal *Biometrika* and published in 1908.”

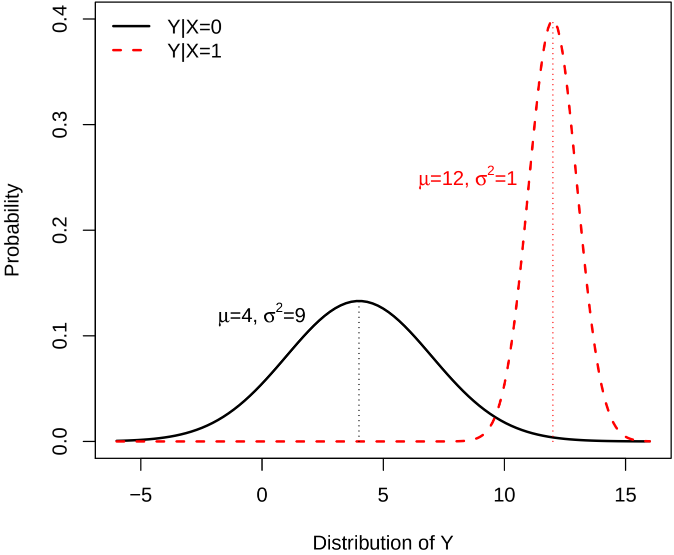
- Student’s t-test (Wikipedia)

The Setup

We have:

- N observations, $i \in \{1, 2, ...N\}$
- A dichotomous predictor X, so that $X_i \in \{0, 1\}$
- n_0 and n_1 are the number of observations in the data with $X = 0$ and $X = 1$, respectively (so $n_0 + n_1 = N$)
- A continuous (interval/ratio) outcome variable Y, with
  - $Y|X = 0 \sim N(\mu_0, \sigma_0^2)$ and
  - $Y|X = 1 \sim N(\mu_1, \sigma_1^2)$.
- Call
  - $\bar{Y}_0 = \bar{Y}|X = 0$, and
  - $\bar{Y}_1 = \bar{Y}|X = 1$

Example



Y|X=0
Y|X=1
μ=12, σ²=1
μ=4, σ²=9
Probability
Distribution of Y
0.0
0.1
0.2
0.3
0.4
−5
0
5
10
15

Difference of (sample) means:

$$\bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i}$$

Has:

$$E(\bar{Y}_1 - \bar{Y}_0) = \mu_1 - \mu_0$$

and

$$Var(\bar{Y}_1 - \bar{Y}_0) = \sigma^2_{\mu_1 - \mu_0}.$$

Difference of Means (continued)

We can show that:

$$\sigma^2_{\mu_1 - \mu_0} = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

In practice we use:

$$s^2_{\bar{Y}_1 - \bar{Y}_0} = \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}$$

The t Statistic

The t-statistic is:

$$
\begin{aligned}
t &= \frac{\bar{Y}_1 - \bar{Y}_0}{s_{\bar{Y}_1 - \bar{Y}_0}} \\
&= \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}}
\end{aligned}
$$

We can show that:

$$t \sim t(\nu)$$

where ν ("nu") is the *degrees of freedom* of the t distribution:

$$\nu \approx \frac{\left(\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1} \right)^2}{\frac{s_0^4}{n_0^2(n_0-1)} + \frac{s_1^4}{n_1^2(n_1-1)}}$$

Other Uses

Test statistic for $H_0: \mu_1 - \mu_0 = k$ is:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_0) - k}{s_{\bar{Y}_1 - \bar{Y}_0}}$$

The $(1 - \alpha) \times 100$ c.i. for $\bar{Y}_1 - \bar{Y}_0$ is:

$$(\bar{Y}_1 - \bar{Y}_0) \pm t_{\alpha/2}(s_{\bar{Y}_1 - \bar{Y}_0}),$$

Differences of Proportions

For a proportion:

$$E(\mu) = \pi$$

and

$$\sigma_{\mu}^{2} = \frac{\pi(1-\pi)}{\mathfrak{N}}.$$

So $\hat{\pi} = \bar{Y}$ and:

$$\begin{aligned} s^2 &= \frac{\hat{\pi}(1-\hat{\pi})}{N} \\ &= \frac{\bar{Y}(1-\bar{Y})}{N}, \end{aligned}$$

For two samples with $\bar{Y}_0$ and $\bar{Y}_1$:

$$s_0 = \sqrt{\frac{\bar{Y}_0(1-\bar{Y}_0)}{n_0}} \quad \text{and} \quad s_1 = \sqrt{\frac{\bar{Y}_1(1-\bar{Y}_1)}{n_1}}$$

Differences of Proportions (cont'd)

This means that

$$z = \frac{(\bar{Y}_0 - \bar{Y}_1)}{\sqrt{\frac{\bar{Y}_0(1-\bar{Y}_0)}{n_0}} + \sqrt{\frac{\bar{Y}_1(1-\bar{Y}_1)}{n_1}}}$$

is $\sim N(0,1)$ for whether the two proportions are different from each other.

Note also that:

$$z^2 \sim \chi^2_1,$$

which (as we'll see next week) is equivalent to a chi-square test for the independence of two variables in a 2×2 table.

Two-Sample t-test

Key things to remember:

- Assumes $Y \sim \; i.i.d. \; N(\mu, \sigma^2)$
  - *Independence* (vs. dependence)
  - *Normality* (vs. skewness)
- Note that if $s_0^2 = s_1^2$, then $\nu = n_0 + n_1 - 2$.
- $\nu = n_0 + n_1 - 2$ is also appropriate if n_0 and $n_1 > 50$ or so

Variances, Independence, & Skewness

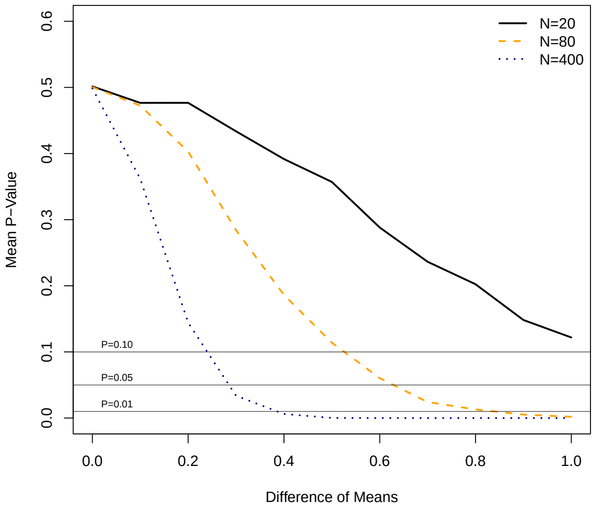
A simulation:

- $Y_0 \sim N(0, 1)$
- $Y_1 \sim N(\mu_1, \sigma_1^2)$
- $\mu_1 \in \{0, 0.1, 0.2, ...1.0\}$
- $\sigma_1^2 \in \{1, 25\}$
- $Y_0, Y_1 \in \{\text{independent}, \text{dependent}\}$
- $N \in \{10, 40, 200\}$ (*per group*)
- $N_{sims} = 1000$

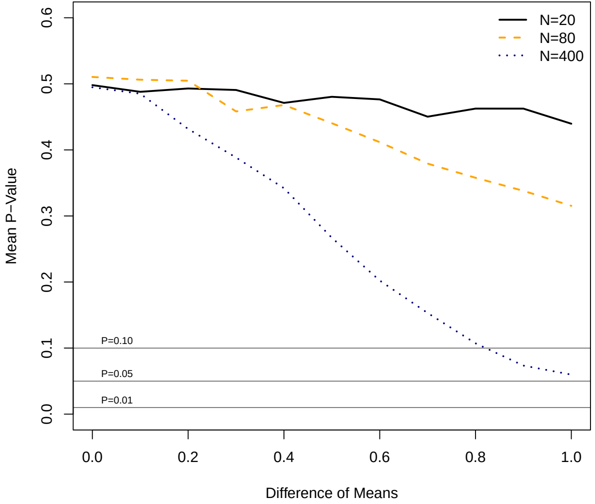
Simulation

```
Nsims <- 1000    # number of sims

N <- c(10,40,200)   # sample sizes
D <- seq(0,1,by=0.1) # differences in means

P1<-as.data.frame(matrix(Nsims,length(N)*length(D)))
set.seed(7222009)

z=1                          # counter...
for(j in 1:length(N)){
  for(k in 1:length(D)){
    for(i in 1:Nsims){
      x<-rnorm(N[j],0,1)             # independent samples,
      y<-rnorm(N[j],0+D[k],1)        # same variance...
      t<-t.test(x,y,var.equal=TRUE) # t-test
      P1[i,z]<-t$p.value             # P-value
    }
    z<-z+1                   # increment counter
  }
}
```

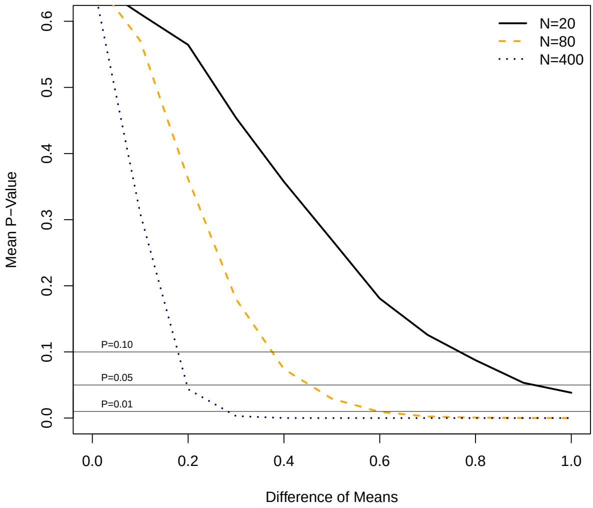
Equal Variances, Independent Samples



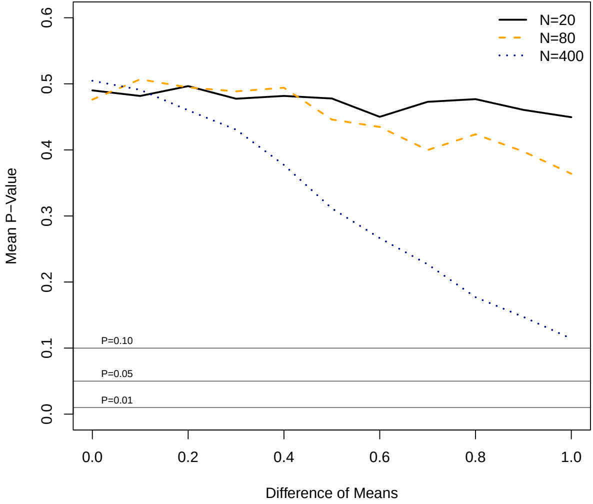
Different Variances, Independent Samples



Equal Variances, Dependent Samples



High Skewness ($Y \sim$ Exponential with $\lambda = 4$)



t-Test Mnemonics

Rough Values of t You'll Want To Get To Know

Absolute Value of t	One-Tailed P-Value*	Two-Tailed P-Value
≈ 1.3	0.10	0.20
≈ 1.65	0.05	0.10
≈ 2	0.025	0.05
≈ 2.4	0.01	0.02
≈ 2.6	0.005	0.01
> 3	< 0.001	< 0.002

Note: All assume d.f. $= \infty$. * indicates that the directionality of the difference in means is "correct."

Example: Federal District Court Judges

The Biographical Directory of Article III Federal Judges contains "the biographies of judges presidentially appointed to serve during good behavior since 1789 on the U.S. district courts, U.S. courts of appeals, Supreme Court of the United States, and U.S. Court of International Trade, as well as the former U.S. circuit courts, Court of Claims, U.S. Customs Court, and U.S. Court of Customs and Patent Appeals."

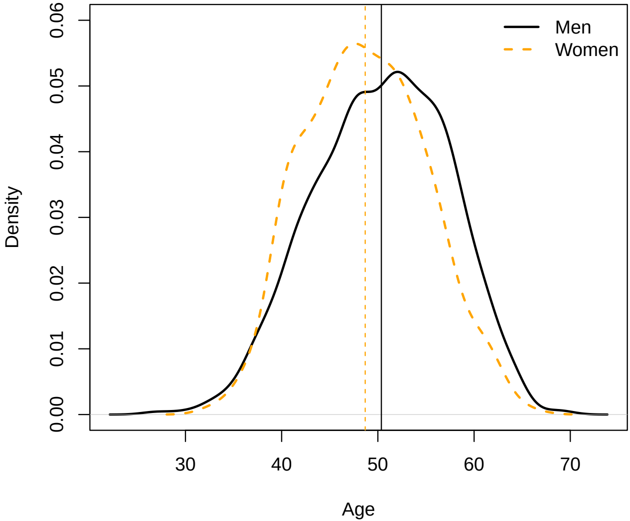
Here: Federal district court judges:

- First appointments *only*
- $N = 3194$ (as of Friday; 3167 after missing data removed)
- Variables of interest:
  - `AppAge`: The age at which each judge was appointed
  - `Gender`: The sex (male or female) of the appointee

Federal District Court Judges

```
> describe(Js$AppAge)
   vars    n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 3167 50.12 6.87     50   50.22 7.41  26  70    44 -0.15    -0.41 0.12


> table(Js$Gender)

Female   Male
   454   2740


> tapply(Js$AppAge,Js$Gender,describe) # Appointment age by gender

$Female
   vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 451 48.68 6.22     48   48.57 7.41  33  66    33 0.12    -0.54 0.29

$Male
   vars    n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 2716 50.36 6.95     51    50.5 7.41  26  70    44 -0.2    -0.37 0.13
```

D. Ct. Judge Appointment Age by Gender



$$\bar{Y}_{Male} - \bar{Y}_{Female} = 1.68$$

and

$$\begin{aligned} s^2_{\bar{Y}_{Male}-\bar{Y}_{Female}} &= \frac{s^2_{Male}}{n_{Male}} + \frac{s^2_{Female}}{n_{Female}} \\ &= \frac{48.30}{2716} + \frac{38.69}{451} \\ &= 0.0178 + 0.0858 \\ &= 0.1036 \end{aligned}$$

and:

$$\begin{aligned} s_{\bar{Y}_{Male}-\bar{Y}_{Female}} &= \sqrt{0.1036} \\ &= 0.322 \end{aligned}$$

Then:

$$\begin{aligned} t &= \frac{1.69 - 0}{0.3220} \\ &= \mathbf{5.22} \end{aligned}$$

t-test (via `t.test`)

```
> T1<-t.test(AppAge~Gender,data=Js)
> T1

Welch Two Sample t-test

data:  AppAge by Gender
t = -5.2, df = 651, p-value = 0.0000002
alternative hypothesis: true difference in means between group Female
  and group Male is not equal to 0
95 percent confidence interval:
 -2.312 -1.048
sample estimates:
mean in group Female   mean in group Male
               48.68                50.36
```

“Reverse” the Difference

```
> Js$Female<-ifelse(Js$Gender=="Female",1,0)

> Ta<-t.test(AppAge~Female,data=Js)
> Ta

Welch Two Sample t-test

data:  AppAge by Female
t = 5.2, df = 651, p-value = 0.0000002
alternative hypothesis: true difference in means between group 0
  and group 1 is not equal to 0
95 percent confidence interval:
 1.048 2.312
sample estimates:
mean in group 0 mean in group 1
          50.36           48.68
```

$H_0: \overline{\text{AppAge}}_{Male} > \overline{\text{AppAge}}_{Female}$

```
> Tg<-t.test(AppAge~Female,data=Js,alternative="greater")
> Tg

Welch Two Sample t-test

data:  AppAge by Female
t = 5.2, df = 651, p-value = 0.0000001
alternative hypothesis: true difference in means between group 0
  and group 1 is greater than 0
95 percent confidence interval:
 1.149   Inf
sample estimates:
mean in group 0 mean in group 1
          50.36           48.68
```

$$H_0: \overline{\text{AppAge}}_{Male} < \overline{\text{AppAge}}_{Female}$$

```
> Tl<-t.test(AppAge~Female,data=Js,alternative="less")
> Tl

Welch Two Sample t-test

data:  AppAge by Female
t = 5.2, df = 651, p-value = 1
alternative hypothesis: true difference in means between group 0
  and group 1 is less than 0
95 percent confidence interval:
 -Inf 2.21
sample estimates:
mean in group 0 mean in group 1
          50.36           48.68
```

H_0 : One Year Age Difference

```
> T1<-t.test(AppAge~Female,data=Js,mu=1)
> T1

Welch Two Sample t-test

data:  AppAge by Female
t = 2.1, df = 651, p-value = 0.04
alternative hypothesis: true difference in means between group 0
  and group 1 is not equal to 1
95 percent confidence interval:
 1.048 2.312
sample estimates:
mean in group 0 mean in group 1
          50.36           48.68
```

Forcing Equal Variances

```
> Te<-t.test(AppAge~Female,data=Js,var.equal=TRUE)
> Te

Two Sample t-test

data:  AppAge by Female
t = 4.8, df = 3165, p-value = 0.000001
alternative hypothesis: true difference in means between group 0
  and group 1 is not equal to 0
95 percent confidence interval:
 0.9968 2.3627
sample estimates:
mean in group 0 mean in group 1
          50.36           48.68
```

Differences of Proportions

Compare white / non-white with gender:

```
> Js$NonWhite<-ifelse(Js$`Race or Ethnicity`=="White",0,1)

> prop.table(table(Js$NonWhite))

     0      1
0.8647 0.1353

> xtabs(~NonWhite+Female,data=Js)
        Female
NonWhite    0    1
       0 2452  310
       1  288  144

> prop.table(xtabs(~NonWhite+Female,data=Js),margin=1)
        Female
NonWhite      0      1
       0 0.8878 0.1122
       1 0.6667 0.3333

> prop.table(xtabs(~NonWhite+Female,data=Js),margin=2)
        Female
NonWhite      0      1
       0 0.8949 0.6828
       1 0.1051 0.3172
```

Difference of Proportions ("by hand")

Is the proportion of female judges the same for white and non-white appointees?

```
> PF<-prop.table(table(Js$Female))[2] # total prop. female
> PFW<-prop.table(xtabs(~NonWhite+Female,data=Js),margin=1)[3] # P(female|white)
> PFNW<-prop.table(xtabs(~NonWhite+Female,data=Js),margin=1)[4]# P(female|nonwhite)
> NM<-table(Js$Female)[1] # N male
> NF<-table(Js$Female)[2] # N female
> s<-sqrt((PF*(1-PF))*((1/NM)+(1/NF))) # s

> Z <- (PFW-PFNW) / s
> Z
    1
-12.5

> pnorm(Z)
        1
3.958e-36

> Z^2    # z-squared is chi-square (1)
    1
156.1

> pchisq(Z^2,df=1,lower.tail=FALSE)
        1
7.916e-36
```

Difference of Proportions

Is the proportion of female judges the same for white and non-white appointees?

```
> Nf<-xtabs(~Js$NonWhite+Js$Female)[c(3,4)]
> Nt<-as.numeric(table(Js$NonWhite))
> FT<-prop.test(Nf,Nt,correct=FALSE)
> FT

2-sample test for equality of proportions without continuity correction

data:  Nf out of Nt
X-squared = 150, df = 1, p-value <2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2671 -0.1751
sample estimates:
prop 1 prop 2
0.1122 0.3333
```

Difference of Proportions (continued)

Is the proportion of non-white judges the same for male and female appointees?

```
> Nf<-xtabs(~Js$NonWhite+Js$Female)[c(3,4)]
> Nt<-as.numeric(table(Js$NonWhite))
> FT<-prop.test(Nf,Nt,correct=FALSE)
> FT

2-sample test for equality of proportions without continuity correction

data:  Nf out of Nt
X-squared = 150, df = 1, p-value <2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2671 -0.1751
sample estimates:
prop 1 prop 2
0.1122 0.3333
```

What We Say

"Among all federal district court judges, the average male judge was 1.68 years older upon appointment than the average female judge ($t = 5.22$, $P < 0.001$)."

and

"Non-white federal district court judges were approximately three times more likely to be female than white judges (33 percent vs. 11 percent, $\chi^2_1 = 150$, $P < 0.001$)."

Power

Power!

Four interrelated components:

- Sample size (N)
- "Effect size" (d)
- Significance level (P):
  - Pr(finding an effect that is not there) / Pr("false positive")
  - Also written as α
- **Power** ($\mathfrak{P}$):
  - Pr(finding an effect that *is* there) / Pr("true positive")
  - Sometimes written $1 - \beta$

Given any three of these, we can determine the fourth.

What's An "Effect Size"?

The size of an effect – e.g., the difference between $\bar{Y}_0$ and $\bar{Y}_1$ – *depends on the "scale" of Y*.

Solution? Cohen's d:

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

- The *standardized* difference between two means...
- σ is the *pooled standard deviation*:

$$\sigma = \sqrt{\frac{(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2}{n_0 + n_1 - 2}}$$

where $s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$ denotes the variance of Y in each group $\{0 \text{ or } 1\}$.

What's a big value for d?

d	Effect Size
0.01	Teeeeny
0.20	Small
0.50	Medium
1.00	Large
2.00	Huuuuge

Example: The t-test

For a given effect size d and sample size N, the t-statistic for testing the hypothesis $d = 0$ (that is, $\mu_0 = \mu_1$) against the alternative hypothesis $d > 0$ (equivalently, $\mu_0 < \mu_1$) is:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_0) - 0}{s_{\bar{Y}_1 - \bar{Y}_0}}.$$

At $P = 0.05$, we reject $d = 0$ if

$$t > 1.64$$

and if N is large, then $t \to N(0, 1)$, and so we can use a z-statistic instead.

Example (continued)

Now suppose $d > 0$ (and N is still large).

The power $\mathfrak{P}(d)$ of t to detect this fact at $P = 0.05$ is:

$$
\begin{aligned}
\mathfrak{P}(d) &= \Pr(t > 1.64 | d) \\
&= \Pr\left[\frac{(\bar{Y}_1 - \bar{Y}_0) - d + d}{s_{\bar{Y}_1 - \bar{Y}_0}} > 1.64\right] \\
&= \Pr\left[\frac{(\bar{Y}_1 - \bar{Y}_0) - d}{s_{\bar{Y}_1 - \bar{Y}_0}} > \left(1.64 - \frac{d}{s_{\bar{Y}_1 - \bar{Y}_0}}\right)\right] \\
&= 1 - \Pr\left[\frac{(\bar{Y}_1 - \bar{Y}_0) - d}{s_{\bar{Y}_1 - \bar{Y}_0}} < \left(1.64 - \frac{d}{s_{\bar{Y}_1 - \bar{Y}_0}}\right)\right] \\
&\approx 1 - \Phi\left(1.64 - \frac{d}{s_{\bar{Y}_1 - \bar{Y}_0}}\right)
\end{aligned}
$$

What's the Point?

So:

$$\mathfrak{P}(d) \approx 1 - \Phi\left(1.64 - \frac{d}{s_{\bar{Y}_1 - \bar{Y}_0}}\right)$$

- Power increases as d gets larger...
- For a given value of d, bigger $N \rightarrow$ higher power (via $s_{\bar{Y}_1 - \bar{Y}_0}$)...
- For very small values of d, power will be low
  - The minimum value of $\mathfrak{P}(d)$ as $d \rightarrow 0$ is P
  - For very small values of d, the difference between $d = 0$ and $d > 0$ is usually unimportant

Hypothetical Example

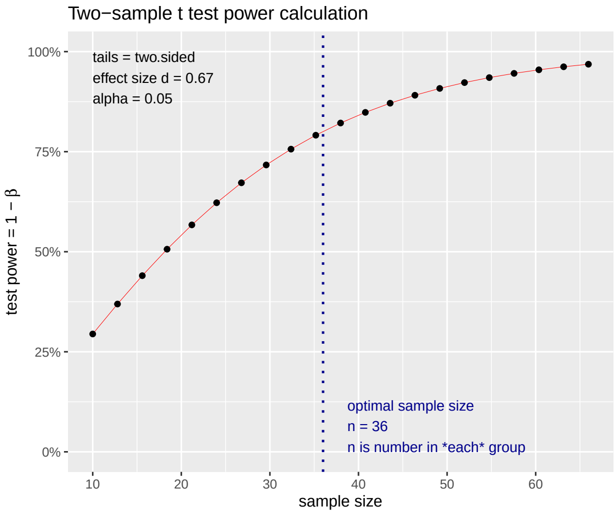
Consider a survey with a standard 101-point "feeling thermometer" (FT) for President Biden. You want to be able to detect the presence of (at the minimum) a 20-point difference in that 101-point scale (say, between Democrats and Republicans) with 80 percent power [$\mathfrak{P} = 0.80$] at $P = 0.05$ (two-tailed). **How big does your sample N need to be?**

Suppose:

- $\sigma_{FT} = 30$, which means
- $d = \frac{20}{30} = 0.67...$

```
> pwr.t.test(d=0.67,sig.level=0.05,power=0.80)

     Two-sample t test power calculation

              n = 35.96
              d = 0.67
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Sample Size Plot



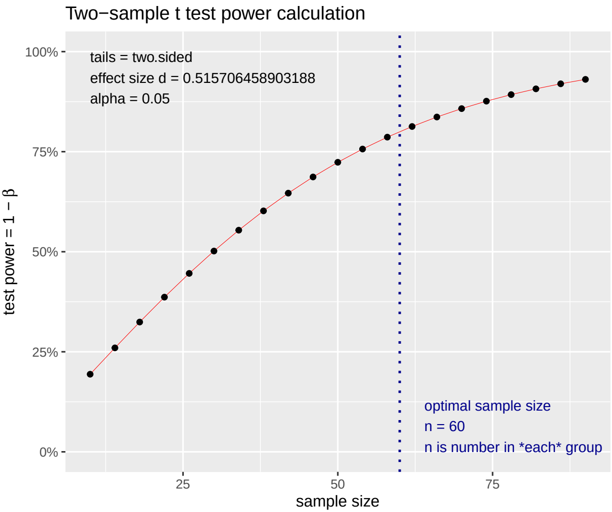
Another Example

I have a small survey with $N = 120$ (total). Given that same 101-point "feeling thermometer," what is the smallest difference d I can detect with $\mathfrak{P} = 0.80$ and $P = 0.05$ (two-tailed)?

```
> pwr.t.test(n=60,sig.level=0.05,power=0.80)

     Two-sample t test power calculation

              n = 60
              d = 0.5157
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Note that here, if $\sigma_{FT} = 30$, the actual size of the smallest difference we can detect with $\mathfrak{P} = 0.80$ and $P = 0.05$ is $(0.5157 \times 30) \approx 15.5$ units on the "raw" feeling thermometer scale.

Effect Size Plot



Conducting Power Analyses

How?

- Lots of power calculators on the internet...
- In R, the `pwr` package:
  - Power calculations for t-tests + many others
  - Can specify tailedness, other options
  - Semi-nice plots

Practical considerations:

- Prospective, and largely geared towards experiments (where N is controlled)
- Requires knowledge of d, which we often don't have...
- We (in political science) don't do this enough; BUT
- Retrospective / post-hoc power analyses are bad