Confidence Intervals and Hypothesis Testing

Foundations of Statistical Inference (PLSC 503)

Evaluating Estimators (Review)

- $\hat{\theta}_n$ is an **unbiased** estimator for θ if $\mathbb{E}[\hat{\theta}_n] = \theta$.
  - i.e. $\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$
- $\hat{\theta}_n$ is a **consistent** estimator for θ if $\hat{\theta}_n \xrightarrow{P} \theta$
  - i.e. $\Pr(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ as $n \to \infty$ for every $\epsilon > 0$.
- **Mean Squared Error**: $\mathbb{E}[\hat{\theta}_n - \theta]^2 = \text{Var}[\hat{\theta}_n] + \text{bias}^2(\hat{\theta}_n)$

Example:

- Suppose X is some random variable from an unknown distribution with finite moments
- Let $Y = \sin(X) + \sqrt{X} + X^3$

Question: is the sample mean $\bar{Y}$ an unbiased and consistent estimator for $\mathbb{E}[Y]$? What's the MSE?

Data generation process for toy example:

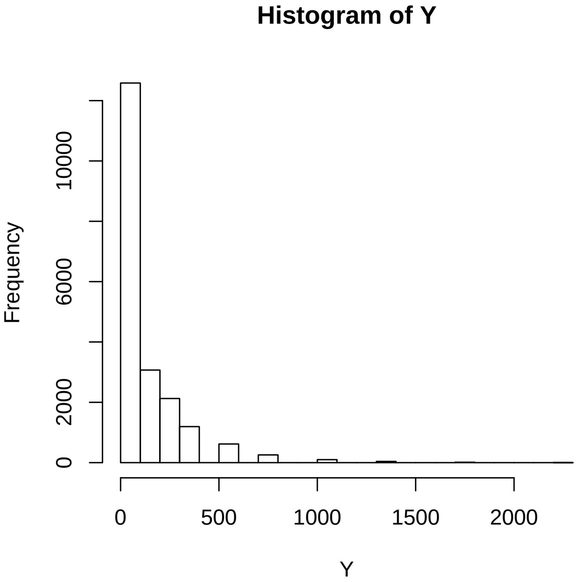
```r
set.seed(503)

# Suppose X ~ Poisson(4)
X <- rpois(n = 20000, lambda = 4)

# Generate Y
Y <- sin(X) + sqrt(X) + X^3

# What's E[Y]?
mean(Y)
```

```
## [1] 117.9372
```


Histogram of Y
Frequency
0
2000
6000
10000
0
500
1000
1500
2000
Y

Evaluating Estimators (Review)

Let $Y_1, \ldots, Y_n$ be i.i.d random draws from Y (with finite $\mathbb{E}[Y]$ and $\text{Var}[Y] > 0$) s.t. $\bar{Y}_n = \frac{Y_1 + \cdots + Y_n}{n}$, $\mathbb{E}[\bar{Y}_n] = \mu$, $\text{Var}[\bar{Y}_n] = \frac{\sigma^2}{n}$.

- $\text{bias}(\bar{Y}_n) = \mathbb{E}[\bar{Y}_n] - \mathbb{E}[Y] = 0$:

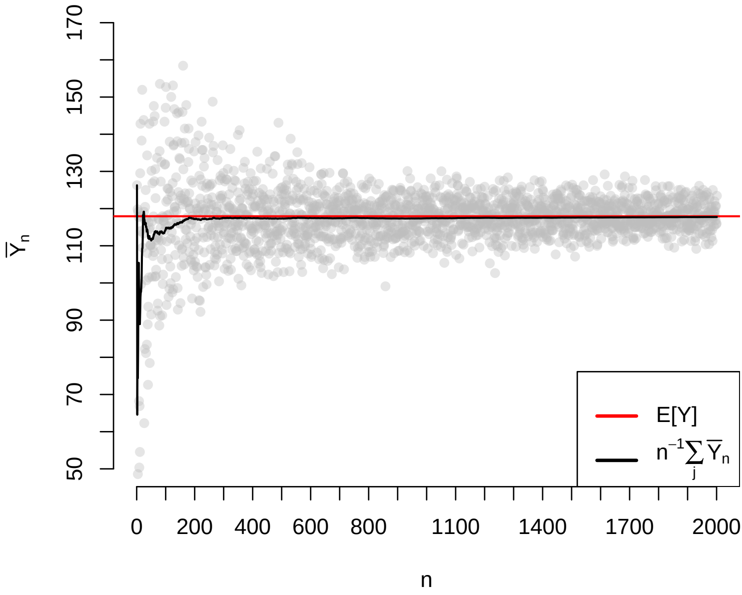
$$\begin{aligned}\mathbb{E}[\bar{Y}_n] =& \mathbb{E}\left[\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)\right] = \frac{1}{n}\mathbb{E}\left[Y_1 + Y_2 + \cdots + Y_n\right] \\ =& \frac{1}{n}\left(\mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \cdots + \mathbb{E}[Y_n]\right) = \frac{1}{n} n \mathbb{E}[Y] = \mathbb{E}[Y]\end{aligned}$$

- $\bar{Y}_n \xrightarrow{p} \mathbb{E}[Y]$: $\Pr\{|\bar{Y}_n - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}$ by Chebyshev's inequality

$$\lim_{n \to \infty} \left(\Pr\{|\bar{Y}_n - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2} \right) = \Pr\{|\bar{Y}_n - \mu| \geq \epsilon\} \leq 0$$

- $\text{MSE}(\bar{Y}_n) = \text{Var}[\bar{Y}_n] = \sigma^2/n \to 0$ as $n \to \infty$

Evaluating Estimators (Review)



Confidence Intervals

- A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$.
- (a, b) traps θ with probability $1-\alpha$, i.e. $\Pr(\theta \in C_n) \geq 1-\alpha$
- C_n is a random variable, θ is fixed!

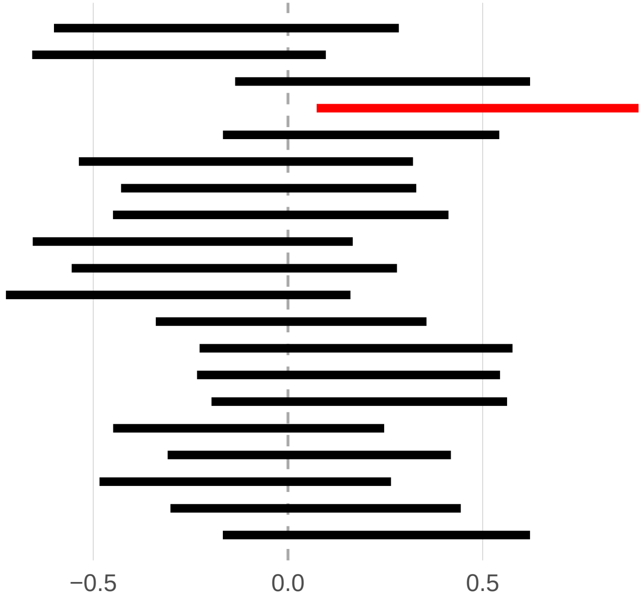
Example: Professor P. Hacker has a novel theory that says shark attacks, on average, affect voting behavior[1]

- He runs 20 experiments on MTurk with 100 subjects each
- Each experiment uses simple random assignment
- The treatment group views a short video about rising shark attacks and then fills out a survey about voting behavior
- P. Hacker's estimator is the diff-in-means, $\hat{\tau}_1, \ldots, \hat{\tau}_{20}$
- Each time, he constructs a valid 95% confidence interval

Question: how many of P. Hacker's intervals will trap τ?

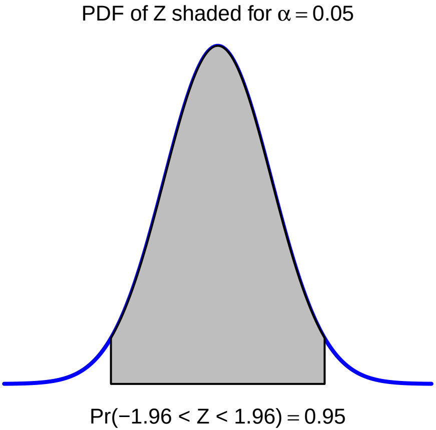
[1] Assume the true Average Treatment Effect (ATE) $\tau = 0$

Professor P. Hacker's Confidence Intervals



Normal-based Confidence Interval

If $Z \sim \mathcal{N}(0,1)$ then $\Pr(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. [2]



[2] where $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$

Normal-based Confidence Interval

- Suppose $(\hat{\theta}_n - \theta)/\hat{\sigma}[\hat{\theta}_n] \xrightarrow{d} \mathcal{N}(0, 1)$
- Let $C_n = \left(\hat{\theta}_n - z_{\alpha/2}\hat{\sigma}[\hat{\theta}_n], \hat{\theta}_n + z_{\alpha/2}\hat{\sigma}[\hat{\theta}_n]\right)$

Question: what is $\Pr(\theta \in C_n)$?

$$\begin{aligned}
\Pr(\theta \in C_n) &= \Pr\left(\hat{\theta}_n - z_{\alpha/2}\hat{\sigma}[\hat{\theta}_n] < \theta < \hat{\theta}_n + z_{\alpha/2}\hat{\sigma}[\hat{\theta}_n]\right) \\
&= \Pr\left(-z_{\alpha/2} < \frac{\hat{\theta}_n - \theta}{\hat{\sigma}[\hat{\theta}_n]} < z_{\alpha/2}\right) \\
&\to \Pr\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1 - \alpha
\end{aligned}$$

Normal-based intervals only have approximate (large sample) coverage guarantees

Normal-based Confidence Interval

Example:

- Recall that $Y = \sin(X) + \sqrt{X} + X^3$.
- We showed $\bar{Y}_n$ was unbiased and consistent for $\mathbb{E}[Y] = \mu$.
- By the Central Limit Theorem:
  - $\bar{Y}_n \overset{A}{\to} \mathcal{N}(\mu, \hat{\sigma}[\bar{Y}_n]^2)$, and $(\bar{Y}_n - \mu)/\hat{\sigma}[\bar{Y}_n] \overset{d}{\to} \mathcal{N}(0, 1)$
- Therefore, $\bar{Y}_n \pm z_{\alpha/2}\hat{\sigma}[\bar{Y}_n]$ is an approximate $1 - \alpha$ CI

```
y <- sample(Y, 1000) # Take 1000 draws from Y
y_bar <- mean(y) # Estimated mean
se_hat <- sd(y)/sqrt(1000) # Estimated SE

# Construct an approximate 89% CI:
c(y_bar - qnorm(1-(0.11/2))*se_hat,
  y_bar + qnorm(1-(0.11/2))*se_hat)
```

```
## [1] 119.5679 138.1325
```

Normal-based Confidence Interval

```
get_ci <- function(alpha = 0.05, n = 1000, dist = Y){
  y <- sample(dist, n)
  y_bar <- mean(y)
  se_hat <- sd(y)/sqrt(n)
  c(y_bar - qnorm(1-(alpha/2))*se_hat,
    y_bar + qnorm(1-(alpha/2))*se_hat)
}

# Make R = 1000 confidence intervals
R <- 1000
out <- replicate(R, get_ci(alpha = 0.11))

# What proportion cover E[Y]?
sum(mean(Y) >= out[1, ] & mean(Y) <= out[2, ])/R
```

```
## [1] 0.902
```

Hypothesis Testing

A **hypothesis test** can be seen as a probabilistic proof by contradiction.

1. Start with some default theory – the **null hypothesis** H_0 – and assume it is true.
2. Pick a test statistic T, which is a function of the **observed data**, e.g. sample mean.
3. Derive the sampling distribution of T when H_0 is true.
4. Calculate the probability of seeing a test statistic as extreme as T^*, **assuming the null is true**.
5. If P is small (i.e. T^* sufficiently unusual) then reject H_0, else retain H_0.

Hypothesis Testing

Example: P. Hacker has conducted 20 experiments. Now he wants to get published, which he suspects is a "coin flip".

1. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$ be n independent submissions. Choose $H_0 : \theta = 0.5$ and $H_1 : \theta \neq 0.5$
2. Let $T = T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$
3. Under H_0, $T \sim \text{Binom}(n, \theta = 0.5)$ and $\mathbb{E}[T] = n0.5$
4. Suppose $T^* = 1$ for $n = 20$, i.e. $|1 - n0.50| = 9$.
$$\begin{aligned} P &= \Pr\left(|T - 10| \geq 9 \mid \theta = 0.5\right) \\ &= \Pr(T \leq 1 \mid \theta = 0.5) + \Pr(T \geq 19 \mid \theta = 0.5) \\ &\approx 4 \times 10^{-5} \end{aligned}$$
5. P is approx. 1 in 25,000, e.g. reject H_0 for $\alpha \gtrapprox 4 \times 10^{-5}$

```
sum(dbinom(0:1, size = 20, prob = 0.5)) +
  sum(dbinom(19:20, size = 20, prob = 0.5))
```

```
## [1] 4.005432e-05
```

Hypothesis Testing

P is **exact** if we know the null distribution. We usually don't...

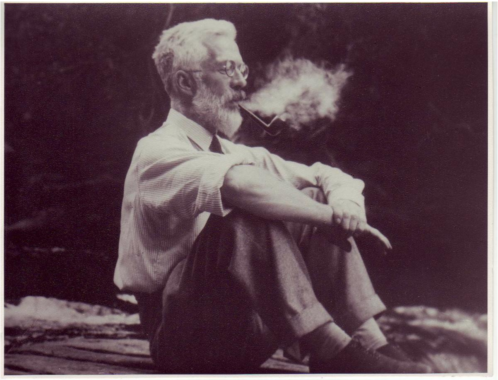
The Wald Test

- $H_0 : \theta = \theta_0$ v.s. $H_1 : \theta \neq \theta_0$ where $(\hat{\theta} - \theta_0)/\hat{\sigma}[\hat{\theta}] \xrightarrow{d} \mathcal{N}(0,1)$
- A size α Wald test: reject H_0 if $|W| > z_{\alpha/2}$
  - for $W = (\hat{\theta} - \theta_0)/\hat{\sigma}[\hat{\theta}]$
- The Wald Test is **asymptotically valid**:

$$
\begin{aligned}
\Pr\left(|W| > z_{\alpha/2}\right) &= \Pr\left(\frac{(\hat{\theta} - \theta_0)}{\hat{\sigma}[\hat{\theta}]} > z_{\alpha/2}\right) \\
&\to \Pr(|Z| > z_{\alpha/2}) \\
&= \alpha
\end{aligned}
$$

Question: what is α? What is $z_{\alpha/2}$?

Hypothesis Testing



"The value for which P=.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."
-Statistical Methods for Research Workers, 1925.

Hypothesis Testing

	Decision	
	Retain H_0	Reject H_0
H_0 true	✓	False Positive
H_0 false	False Negative	✓

- **Significance level** of a test: $\Pr(\text{reject } H_0 \mid H_0 \text{ true})$
  - $\Pr(\text{reject } H_0 \mid H_0 \text{ true}) = 0.05$ for a test with $\alpha = 0.05$.
- **Power** of a test: $\Pr(\text{reject } H_0 \mid H_0 \text{ false})$.
  - Often written $1 - \beta$ where $\beta = \Pr(\text{retain } H_0 \mid H_0 \text{ false})$
- **Minimum Detectable Effect**: $\left(z_{\alpha/2} + z_{\beta}\right)\sigma[\hat{\theta}]$
  - Large sample approximation for two-sided hypothesis testing
  - For $\alpha = 0.05$, $\beta = 0.20$, $\text{MDE} = (1.96 + 0.84)\sigma[\hat{\theta}] = 2.8\sigma[\hat{\theta}]$
  - $\sigma[\hat{\theta}] \to 0$ as $n \to \infty$, so $\text{MDE} \to 0$ as $n \to \infty$

Hypothesis Testing

1. $H_0 : \tau = 0$ v.s. $H_1 : \tau \neq 0$ for $\tau = \mu_t - \mu_c$.
2. Choose $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$ with $\hat{\sigma}[\hat{\tau}] = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$
3. Under H_0, $W = \frac{(\hat{\tau} - 0)}{\hat{\sigma}[\hat{\tau}]} = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}} \xrightarrow{d} \mathcal{N}(0, 1)$
4. Suppose $W^* = \frac{0.49}{\sqrt{\frac{1.07}{50} + \frac{1.02}{50}}} \approx 2.40$
5. Is $P < 0.05$?

$$
\begin{aligned}
P &\approx \Pr(|W| > 2.40 \mid \tau = 0) \\
&= \Pr(W < -2.40 \mid \tau = 0) + \Pr(W > 2.40 \mid \tau = 0) \\
&\approx 0.02
\end{aligned}
$$

```r
pnorm(-2.40) + 1-pnorm(2.40)
```

```
## [1] 0.01639507
```

Statistical Power

Question: Suppose the null is false, e.g. $\tau \neq \tau_0$, and fix $\alpha = 0.05$. What is the **power** of a Wald test?

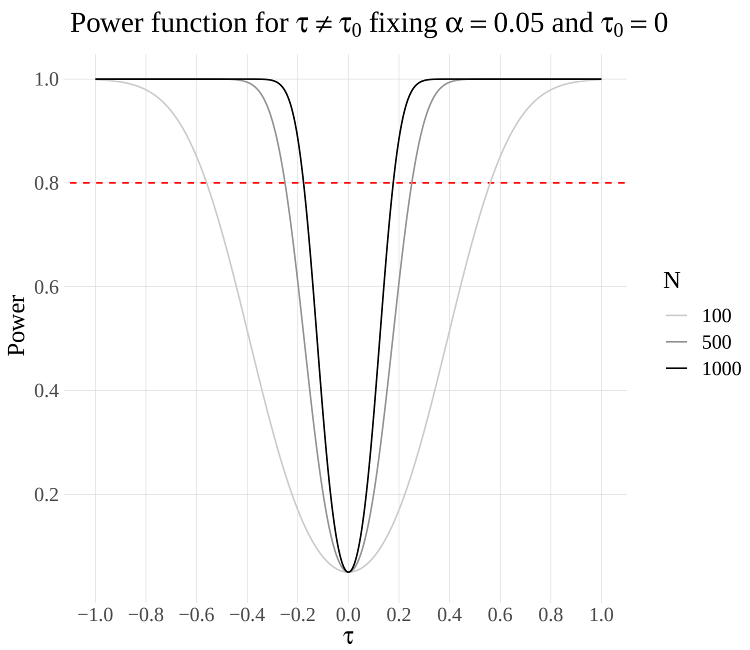
$$
\begin{aligned}
\Pr(\text{reject} \;\; H_0 \mid \tau \neq \tau_0) &= \Pr\left(W < -c \;\; \mid \tau \neq \tau_0\right) + \Pr\left(W > c \;\; \mid \tau \neq \tau_0\right) \\
&= \Pr\left(\frac{\hat{\tau} - \tau_0}{\hat{\sigma}[\hat{\tau}]} < -c \;\; \mid \tau \neq \tau_0\right) + \\
&\qquad \Pr\left(\frac{\hat{\tau} - \tau_0}{\hat{\sigma}[\hat{\tau}]} > c \;\; \mid \tau \neq \tau_0\right) \\
&= \Pr\left(Z < -c + \frac{\tau_0 - \tau}{\hat{\sigma}[\hat{\tau}]}\right) + \Pr\left(Z > c + \frac{\tau_0 - \tau}{\hat{\sigma}[\hat{\tau}]}\right) \\
&= \Phi\left(\frac{\tau_0 - \tau}{\hat{\sigma}[\hat{\tau}]} - c\right) + 1 - \Phi\left(\frac{\tau_0 - \tau}{\hat{\sigma}[\hat{\tau}]} + c\right) \\
&= \Phi\left(-z_{\alpha/2} + \frac{\tau_0 - \tau}{\hat{\sigma}[\hat{\tau}]}\right) + \Phi\left(-z_{\alpha/2} - \frac{\tau_0 - \tau}{\hat{\sigma}[\hat{\tau}]}\right)
\end{aligned}
$$

Statistical Power

```
wald_power <- function(tau = NULL , tau_0 = 0,
                        n_t = NULL, n_c = NULL,
                        s2_t = NULL, s2_c = NULL,
                        a = 0.05){
  z <- qnorm(1-(a/2))
  se_hat <- sqrt(s2_t/n_t + s2_c/n_c)
  pnorm(-z + (tau_0 - tau)/se_hat) +
    pnorm(-z - (tau_0 - tau)/se_hat)
}

# Assume tau = 0.49, s2_t = s2_c = 1
wald_power(tau = 0.49, n_t = 50, n_c = 50,
           s2_t = 1, s2_c = 1)
```

```
## [1] 0.687951
```


Power function for $\tau \neq \tau_0$ fixing $\alpha = 0.05$ and $\tau_0 = 0$
1.0
0.8
0.6
0.4
0.2
Power
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
τ
N
100
500
1000

Minimum Detectable Effect

Assume (without proof)[3]: $\text{Var}[\hat{\tau}] \leq \frac{1}{N}\left(\frac{s_t^2}{p} + \frac{s_c^2}{1-p}\right)$

Probability of treatment p, sample variance $s_k^2 = \frac{1}{n_k-1}\sum(Y_{ki} - \bar{Y}_k)^2$

What is the relationship between $N = n_t + n_c$ and MDE?

$$\begin{aligned} \text{MDE} &= (z_{\alpha/2} + z_\beta)\sigma[\hat{\tau}] \\ &= (z_{\alpha/2} + z_\beta)\sqrt{\frac{1}{N}\left(\frac{s_t^2}{p} + \frac{s_c^2}{1-p}\right)} \end{aligned}$$

Or, re-arranging to get: $N = \frac{(z_{\alpha/2}+z_\beta)^2\left(\frac{s_t^2}{p}+\frac{s_c^2}{1-p}\right)}{\text{MDE}^2}$

This is a reasonable approximation w/ large sample

[3] see Aronow, Green and Lee (2014), "Sharp Bounds on the Variance in Randomized Experiments," *The Annals of Statistics*, for better bounds!

Minimum Detectable Effect

Question: What's the MDE when $N = 100, \alpha = 0.05, \beta = 0.20$?

```
# Many moving parts:
get_mde <- function(a = 0.05, b = 0.20, p = 0.5,
                    n = NULL, s2_t = NULL,
                    s2_c = NULL){
  (qnorm(1-(a/2))+ qnorm(1-b))*
    sqrt(n^-1*(s2_t/p + s2_c/(1-p)))
}

get_mde(n = 100, s2_t = 1, s2_c = 1)
```

```
## [1] 0.560317
```

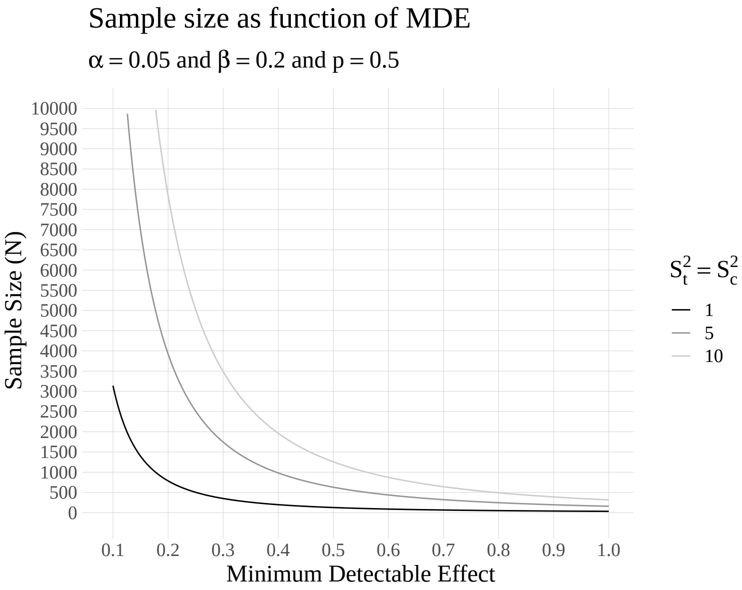
Minimum Detectable Effect

Question: What sample size does he need for an MDE of 0.2 units?

```
get_n <- function(a = 0.05, b = 0.20, p = 0.5,
                  MDE = NULL, s2_t = NULL,
                  s2_c = NULL){
  ((qnorm(1-(a/2))+ qnorm(1-b))^2*
     (s2_t/p + s2_c/(1-p)))/MDE^2
}

get_n(MDE = 0.2, s2_t = 1, s2_c = 1)
```

```
## [1] 784.888
```



Sample size as function of MDE
α = 0.05 and β = 0.2 and p = 0.5
Sample Size (N)
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Minimum Detectable Effect
$S_t^2 = S_c^2$
1
5
10

Multiple Testing

Example: Prof. Dr. P. Hacker has conducted 20 MTurk experiments to test his novel theory about shark attacks.

- P. Hacker observes one of his 20 P-values is 0.02, which corresponds to an estimated effect of 0.49 units!
- He reports this result, arguing "If shark attacks **do not** affect voting behavior, the probability of observing a result as extreme as 0.49 is 0.02; so they must!"
- P. Hacker, *forthcoming*, "Shark Attacks Affect Voting Behavior: $P < 0.05$", *American Political Science Review*[4]

Question: What is the probability of at least one $P < 0.05$?

$$\begin{aligned}\Pr(\text{at least one significant result}) &= 1 - \Pr(\text{no significant results}) \\ &= 1 - (1 - 0.05)^{20} \\ &\approx 0.64\end{aligned}$$

[4] cf. J. Cohen, 1994, "The earth is round ($p < .05$)," *American Psychologist*

Multiple Testing

Example: Prof. Dr. P. Hacker downloads the latest version of the ANES survey, which contains a question about exposure to shark attacks (z), along with an outcome variable about voting (y).

- He opens Stata and types `reg y x, robust` but $P > 0.05$ ☹
- Suppose the ANES survey has $k = 20$ other predictors

Question: How many possible models can he fit with $k = 20$ predictors?

- Suppose $k = 3$. How many models can he fit?
- $y \sim z$; $y \sim z + x_1 + x_2 + x_3$; $y \sim z + x_k$ (3x); $y \sim z + x_k + x_j$ (3x)
- With $k = 20$ he can fit $2^{20} > 10^6$ simple models
- This yields $10^6 \cdot 0.05 \approx 50000$ "significant" results

Multiple Testing

Potential remedies:

1. Bonferroni correction: reject if $P < \frac{\alpha}{m}$, i.e. $\frac{0.05}{20} = 0.0025$.
   - Ensures probability of false rejection $\leq \alpha$ for any null
   - Easy to implement: just multiply P by m! Very conservative.
2. Control FDR $\leq \frac{m_0}{m}\alpha \leq \alpha$ (e.g. Benjamini-Hochberg)
   - Order $P_{(1)} < \cdots < P_{(m)}$ and find largest P_k s.t. $P_k \leq \frac{k}{m}\alpha$.
   - For $P = (0.01, 0.04, 0.24, 0.58)$, $P_1 < \frac{1}{4}0.05 = 0.0125$, but $P_i > \frac{i}{4}0.05$ for $i \in \{2, 3, 4\}$
3. Pre-registration

See `p.adjust()` in `R`, and Alex Coppock's EGAP guide: 10 Things to Know About Multiple Comparisons