MA678 Homework 3

ZHIXUAN GUAN

9/30/2025

4.4 Designing an experiment

You want to gather data to determine which of two students is a better basketball shooter. You plan to have each student take N shots and then compare their shooting percentages. Roughly how large does N have to be for you to have a good chance of distinguishing a 30% shooter from a 40% shooter? Tell the two shooters apart if one is 30% and the other 40%. Use a two-sided 5% test with 80% power. $\frac{\Delta}{SE} \approx 2.8$

With N shots per student and independent binomial variability,
$\hat{p}_2 - \hat{p}_1$ has
$SE = \sqrt{\frac{p_1(1-p_1)}{N} + \frac{p_2(1-p_2)}{N}}$
Take $p_1 = 0.30$, $p_2 = 0.40$:
$\Delta = 0.10$,
$SE = \sqrt{\frac{0.30 \cdot 0.70 + 0.40 \cdot 0.60}{N}}$
$= \sqrt{\frac{0.45}{N}}$
$= 0.6708/\sqrt{N}$ Set the signal-to-noise ratio to about 2.8:
$\frac{\Delta}{SE} = 2.8$

$$\frac{0.10}{0.6708/\sqrt{N}} = 2.8 \;\Rightarrow\; \sqrt{N} = \frac{2.8 \times 0.6708}{0.10} \approx 18.78 \;\Rightarrow\; N \approx 353$$

So each shooter needs about **360 shots** for an 80% chance to distinguish a 30% shooter from a 40% shooter.

```
pp <- power.prop.test(p1 = 0.30, p2 = 0.40,power = 0.80, sig.level = 0.05,)
ceiling(pp$n)
```

```
## [1] 356
```

```
#So each shooter needs about **360 shots** for an 80% chance to distinguish a 30% shooter from a 40% sh
```

4.6 Hypothesis testing

The following are the proportions of girl births in Vienna for each month in girl births 1908 and 1909 (out of an average of 3900 births per month):

```
birthdata <- c(.4777,.4875,.4859,.4754,.4874,.4864,.4813,.4787,.4895,.4797,.4876,.4859,
               .4857,.4907,.5010,.4903,.4860,.4911,.4871,.4725,.4822,.4870,.4823,.4973)
```

The data are in the folder `Girls`. These proportions were used by von Mises (1957) to support a claim that that the sex ratios were less variable than would be expected under the binomial distribution. We think von Mises was mistaken in that he did not account for the possibility that this discrepancy could arise just by chance.

(a)

Compute the standard deviation of these proportions and compare to the standard deviation that would be expected if the sexes of babies were independently decided with a constant probability over the 24-month period.

```
sd_data<- sd(birthdata)
p <- mean(birthdata)
n<- 3900
sd_exp<- sqrt(p*(1-p)/n)
sd_data
```

```
## [1] 0.006409724
```

```
sd_exp
```

```
## [1] 0.008003121
```

```
#The observed value is slightly smaller than expected, which suggests the data look a bit "too stable,"
```

(b)

The observed standard deviation of the 24 proportions will not be identical to its theoretical expectation. In this case, is this difference small enough to be explained by random variation? Under the randomness model, the actual variance should have a distribution with expected value equal to the theoretical variance, and proportional to a χ^2 random variable with 23 degrees of freedom; see page 53.

```
var_obs    <- sd_data^2
var_theory <- sd_exp^2
df <- n - 1
chi2_stat <- df * var_obs / var_theory
p_two_sided <- 2 * min(pchisq(chi2_stat, df = df),
                       1 - pchisq(chi2_stat, df = df))

c(chi2_stat = chi2_stat, df = df, p_two_sided = p_two_sided)
```

```
##    chi2_stat           df  p_two_sided
## 2.500997e+03 3.899000e+03 2.148221e-74
```

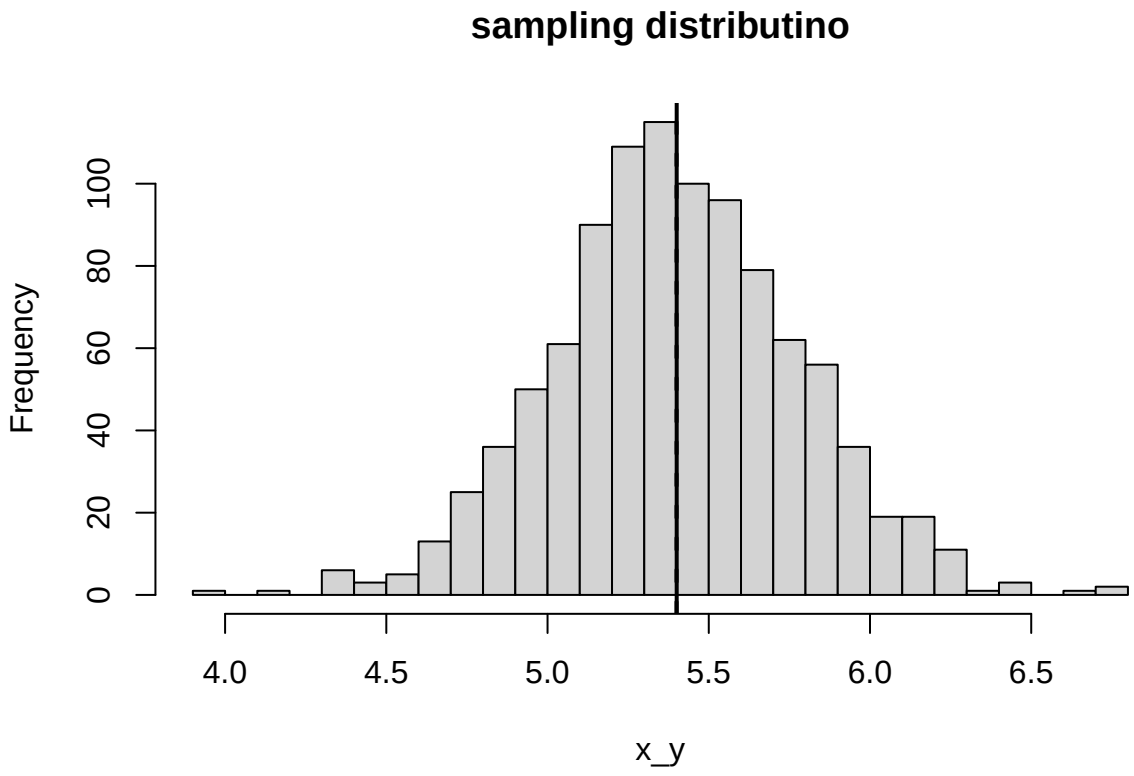
```
#test statistic is tiny (0.09) compared to a chi-square distribution with 23 df.The p-value is extremel
#This means the observed variance is much smaller than what would be expected by chance under the null
#With such a small p-value, we reject the null model: the data show less variability in monthly girl-bi
#This supports von Mises' original observation: sex ratios looked "too stable."
```

5.5 Distribution of averages and differences

The heights of men in the United States are approximately normally distributed with mean 69.1 inches and standard deviation 2.9 inches. The heights of women are approximately normally distributed with mean 63.7 inches and standard deviation 2.7 inches. Let x be the average height of 100 randomly sampled men, and y be the average height of 100 randomly sampled women. In R, create 1000 simulations of $x - y$ and plot their histogram. Using the simulations, compute the mean and standard deviation of the distribution of $x - y$ and compare to their exact values.

```
set.seed(123)
n_m <- 100
n_w<- 100
me_m<- 69.1
me_w <-63.7
```

```r
sd_m<- 2.9
sd_w<- 2.7
x_y <- replicate(1000,mean(rnorm(n_m,me_m,sd_m))-mean(rnorm(n_w,me_w,sd_w)))
mean_stimulation<- mean(x_y)
mean_exact<- me_m-me_w
sd_x_y<- sd(x_y)
sd_exact<- sqrt(sd_m^2/n_m+sd_w^2/n_w)
hist(x_y,breaks = 30, main="sampling distributino")
abline(v = mean_stimulation,  lwd = 2)
abline(v = mean_exact, lwd = 2, lty = 2)
```




```r
#as you can see in the graph,the solid line and the dashed line are perfectly aligned.
#the simulated average difference in heights is practically identical to the theoretical average differ
```

5.8 Coverage of confidence intervals:

On page 15 there is a discussion of an experimental study of an education-related intervention in Jamaica, in which the point estimate of the treatment effect, on the log scale, was 0.35 with a standard error of 0.17. Suppose the true effect is 0.10—this seems more realistic than the point estimate of 0.35—so that the treatment on average would increase earnings by 0.10 on the log scale. Use simulation to study the statistical properties of this experiment, assuming the standard error is 0.17.

(a)

Simulate 1000 independent replications of the experiment assuming that the point estimate is normally distributed with mean 0.10 and standard deviation 0.17.

```r
set.seed(123)
stimu<- rnorm(1000,0.10,0.17)
summary(stimu)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.377662 -0.006815  0.101566  0.102742  0.212982  0.650977
```

```r
#simulation confirms that the estimator is unbiased (mean  0.10) but has substantial variability (many
```

(b)

For each replication, compute the 95% confidence interval. Check how many of these intervals include the true parameter value.

```r
z<- 1.96
ci_lower <- stimu - z * 0.17
ci_upper <- stimu + z * 0.17
covered <- (ci_lower <= 0.1) & (ci_upper >= 0.1)
coverage_rate <- mean(covered)
c(coverage_rate = coverage_rate, expected_95pct = 0.95)
```

```
##  coverage_rate expected_95pct
##          0.947          0.950
```

```r
#converage rate is close to 0.95
```

(c)

Compute the average and standard deviation of the 1000 point estimates; these represent the mean and standard deviation of the sampling distribution of the estimated treatment effect.

```r
mean_sti<- mean(stimu)
sd_sti<- sd(stimu)
mean_sti
```

```
## [1] 0.1027417
```

```r
sd_sti
```

```
## [1] 0.1685881
```

```r
#Across 1000 experiments, the estimates average out to about the true value (0.10) with variability mat
```

10.3 Checking statistical significance

In this exercise and the next, you will simulate two variables that are statistically independent of each other to see what happens when we run a regression to predict one from the other. Generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing `var1 <- rnorm(1000,0,1)` in R. Generate another variable in the same way (call it `var2`). Run a regression of one variable on the other. Is the slope coefficient "statistically significant"? We do not recommend summarizing regressions in this way, but it can be useful to understand how this works, given that others will do so.

```r
var1 <- rnorm(1000,0,1)
var2 <- rnorm(1000,0,1)
fit<- lm(var2~var1)
summary(fit)
```

```
##
```

```
## Call:
## lm(formula = var2 ~ var1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8387 -0.6307 -0.0302  0.6644  3.4057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02120    0.03097  -0.685    0.494
## var1         0.02568    0.03066   0.838    0.402
##
## Residual standard error: 0.9785 on 998 degrees of freedom
## Multiple R-squared:  0.0007024,  Adjusted R-squared:  -0.0002989
## F-statistic: 0.7015 on 1 and 998 DF,  p-value: 0.4025
```

```
#p value for coefficient of var1 and intercept is way larger than 0.5, which means it's not statistical
```

11.3 Coverage of confidence intervals

Consider the following procedure:

- Set $n = 100$ and draw n continuous values x_i uniformly distributed between 0 and 10. Then simulate data from the model $y_i = a + bx_i + \text{error}_i$, for $i = 1, \ldots, n$, with $a = 2$, $b = 3$, and independent errors from a normal distribution.
- Regress y on x. Look at the median and mad sd of b. Check to see if the interval formed by the median $\pm$ 2 mad sd includes the true value, $b = 3$.
- Repeat the above two steps 1000 times.

```
set.seed(123)

B <- 1000
n <- 100
a <- 2; b_true <- 3
sigma <- 1
keep_bhat <- numeric(B)
covered <- logical(B)

for (i in 1:B) {
  x <- runif(n, 0, 10)
  y <- a + b_true*x + rnorm(n, 0, sigma)
  fit <- lm(y ~ x)
  bhat <- coef(fit)[2]
  se_b <- summary(fit)$coef[2,2]
  keep_bhat[i] <- bhat
  ci <- bhat + c(-1,1)*1.96*se_b
  covered[i] <- (ci[1] <= b_true && b_true <= ci[2])
}
#this is a process of generating 95 ci for each stimulation

mean(covered)
```

```
## [1] 0.951
```

```r
med_b <- median(keep_bhat)
mad_sd <- 1.4826 * mad(keep_bhat, center = med_b)
c(median = med_b,
  mad_sd = mad_sd,
  low = med_b - 2*mad_sd,
  high = med_b + 2*mad_sd,
  contains_true_b = (med_b - 2*mad_sd <= b_true && b_true <= med_b + 2*mad_sd))
```

```
##          median          mad_sd             low            high contains_true_b
##      2.99995084      0.04897077      2.90200930      3.09789237      1.00000000
```

```r
#this is mean and sd of b from stimulation, which has nothing to do with 95 ci
```

(a)

True or false: the interval should contain the true value approximately 950 times. Explain your answer. FALSE: as shown in code(ci <- bhat + c(-1,1)*1.96*se_b) this is how to generate a 95% CI. The interval median $\pm$ 2 $\cdot$ mad_sd computed from the 1000 simulated values is not a per-experiment 95% CI; it's a summary of the sampling distribution. You don't expect it to "cover 950 times" because you only evaluate it once. ### (b) Same as above, except the error distribution is bimodal, not normal. True or false: the interval should contain the true value approximately 950 times. Explain your answer. FALSE:same as above:The pooled median $\pm$ 2 $\cdot$ mad_sd across 1000 estimates is still just a summary of the sampling distribution; it's not a per-dataset CI, so the "950 times" logic still doesn't apply.