ENV 710

z-tests and t-tests

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

Exam 1 coming up Wed., Feb. 22!!!

descriptive statistics

discrete probability/distributions

continuous probability/distributions

inference

↓

one- and two-sample tests

z-test, t-tests, etc., more on hypothesis testing and statistical power

⇣

study design

data transformation

0 – covid cases

In 30 towns, a researcher sampled 100 people, 50 people vaccinated against COVID and 50 unvaccinated people. None of the vaccinated people were hospitalized, but several unvaccinated people from each town were. Therefore, assume the population mean of vaccinated people I 0.

If the null hypothesis is no difference in hospitalizations between vaccinated and unvaccinated people, test the following:

$$H_0 : \mu_{\text{vacc}} = \mu_{\text{unvacc}}$$

$$H_1 : \mu_{\text{vacc}} > \mu_{\text{unvacc}}$$

$$H_2 : \mu_{\text{vacc}} \neq \mu_{\text{unvacc}}$$

simulate the number of hospitalized people in the 30 towns with:

```
set.seed(999)
unvacc <- rpois(n = 30, lambda = 3)
```

what is the critical Z value for H_1?
what is the critical Z value for H_2?

what is the p-value for H_1 and H_2?

0 – covid cases

In 30 towns, a researcher sampled 100 people, 50 people vaccinated against COVID and 50 unvaccinated people. None of the vaccinated people were hospitalized, but several unvaccinated people from each town were. Therefore, assume the population mean of vaccinated people I 0.

If the null hypothesis is no difference in hospitalizations between vaccinated and unvaccinated people, test the following:

$$H_0 : \mu_{\text{vacc}} = \mu_{\text{unvacc}}$$

$$H_1 : \mu_{\text{vacc}} > \mu_{\text{unvacc}}$$

$$H_2 : \mu_{\text{vacc}} \neq \mu_{\text{unvacc}}$$

what is the critical Z value for H_1?

```
Z1 <- qnorm(p=0.95, mean=0, sd=1)
[1] 1.644854
```

what is the critical Z value for H_2?

```
Z2 <- qnorm(p=c(0.025, 0.975), mean=0, sd=1)
[1] -1.959964  1.959964
```

what is the p-value for H_1 and H_2?

```
Z <- (mean(unvacc)-0)/(sd(unvacc)/sqrt(30))
[1] 9.405998

h1 <- pnorm(q=Z, mean=0, sd=1, lower.tail=F)
[1] 2.576944e-21

h2 <- pnorm(q=Z, mean=0, sd=1, lower.tail=F)*2
[1] 5.153888e-21
```

I – summary

Discuss the following

- what is the difference between a z-test and a t-test? when would you use either test?
- what are one- and two-sample tests?
- what are one- and two-sided tests?
- what are the assumptions/conditions of one- and two-sample t-tests?
- what is meant by a 'paired test' and state an example of a paired test in your area of expertise.

2 – salmonella

An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were:

```
0.593 0.142 0.329 0.691 0.231 0.793
          0.519 0.392 0.418
```

Is the mean level of Salmonella greater than 0.3 MPN/g, the accepted level?

1. what are the null and alternative hypotheses?
2. state the type of test that you use, e.g., one-sample, two-sample, one-sided, two-sided, etc.
3. what is your conclusion? is the mean level of Salmonella greater than 0.3 MPN/g?
4. write a sentence that articulates the results of your test

```
salmonella <- c(0.593 0.142 0.329 0.691
0.231 0.793 0.519 0.392 0.418)
```

2 – salmonella

An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were:

```
0.593 0.142 0.329 0.691 0.231 0.793
          0.519 0.392 0.418
```

Is the mean level of Salmonella greater than 0.3 MPN/g, the accepted level?

this is a one-sided, one-sample t-test

```
t.test(salmonella, mu=0.3,
alternative="greater")
```

the mean level of Salmonella in the ice cream samples is significantly greater than 0.3 MPN/g (t = 2.205, df = 8, p = 0.029)

3 – placebo

6 subjects were given a drug (treatment group) and an additional 6 subjects were given a placebo (control group). Their reaction time to a stimulus was measured (in milliseconds). Is the mean reaction time slower (e.g., greater) in the treatment than the control group?

```
cont <- c(95, 94, 99, 96, 95, 96)
treat <- c(101, 98, 100, 97, 99, 99)
```

1. state the type of test that you use, e.g., one-sample, two-sample, one-sided, two-sided, z-test, t-test
2. what are the null and alternative hypotheses for this problem?
3. what are the assumptions of the test? check them…
4. write a sentence that articulates the results of your test

3 – placebo

6 subjects were given a drug (treatment group) and an additional 6 subjects were given a placebo (control group). Their reaction time to a stimulus was measured (in milliseconds). Is the mean reaction time slower (e.g., greater) in the treatment than the control group?

1. two-sample, one-sided t-test
2. H_0: no difference in reaction time between the treatment and control groups
   H_a: mean reaction time is slower in the treatment group than the control group, e.g., $\mu_{treat} > \mu_{control}$
3. assumptions: normality & equal variances

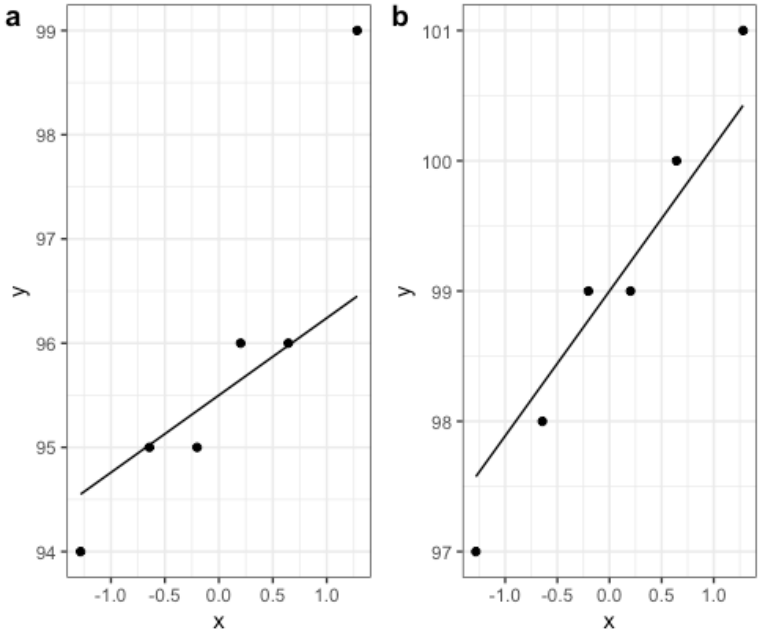
check normality of each sample?

```
> shapiro.test(cont)

Shapiro-Wilk normality test
data:  Cont
W = 0.86587, p-value = 0.2102

> shapiro.test(treat)

Shapiro-Wilk normality test
data:  Treat
W = 0.98176, p-value = 0.96
```

3 – placebo

3. assumptions: normality & equal variances

```
drug <- data.frame(cbind(rxn = c(cont, treat),
         ttt = c(rep("c", length(cont)),
                 rep("t", length(treat)))))
drug$rxn <- as.integer(drug$rxn)

qqc <- ggplot(drug[drug$ttt == "c",], aes(sample = rxn)) +
          stat_qq() + stat_qq_line() +
        theme_bw()

qqt <- ggplot(drug[drug$ttt == "t",], aes(sample = rxn)) +
          stat_qq() + stat_qq_line() +
        theme_bw()

ggpubr::ggarrange(qqc, qqt, nrow=1,
        labels = "auto")
```



3 – placebo

3. assumptions: normality & equal variances

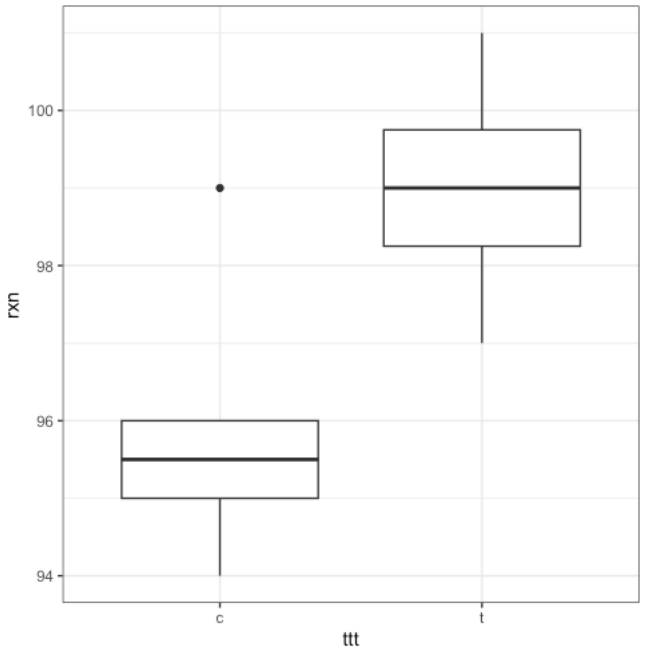
check equal variances of each sample

```
sd(cont)/sd(treat)
```

```
var.test(cont, treat)
```

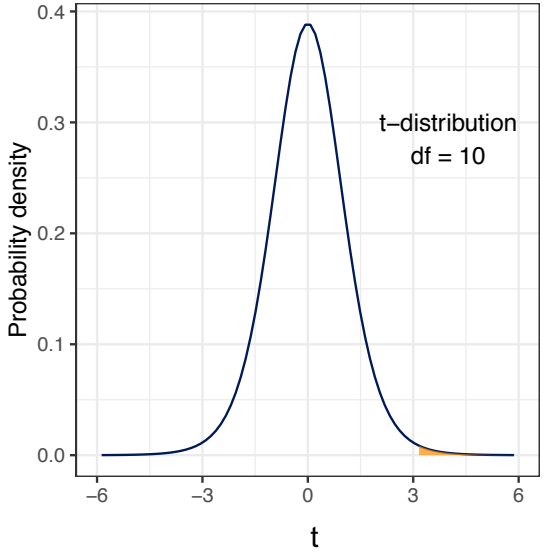
```
        F test to compare two variances

data:  cont and treat
F = 1.4833, num df = 5, denom df = 5,
                p-value = 0.6758
alternative hypothesis: true ratio of variances
              is not equal to 1
        95 percent confidence interval:
             0.2075642 10.6004664
              sample estimates:
             ratio of variances
                       1.483333
```

```
ggplot(drug, aes(y = rxn, x = ttt)) +
      geom_boxplot() +
       theme_bw()
```



3 – placebo

6 subjects were given a drug (treatment group) and an additional 6 subjects were given a placebo (control group). Their reaction time to a stimulus was measured (in milliseconds). Is the mean reaction time slower (e.g., greater) in the treatment than the control group?



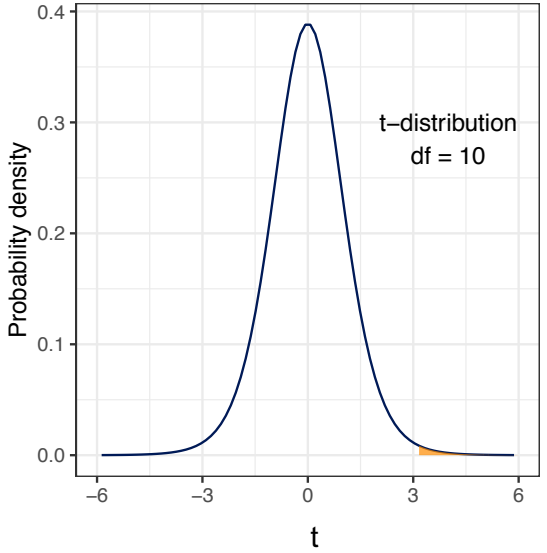

```
        Two Sample t-test
    data:  treat and cont
t = 3.4805, df = 10, p-value = 0.002958
alternative hypothesis: true difference
    in means is greater than 0
  95 percent confidence interval:
     1.517648      Inf
  sample estimates:
mean of x mean of y
 99.00000  95.83333
```

```
t.test(x=treat, y=cont, alternative = "g")
```

alternative = "g"
tests if x has a greater mean than y

3 – placebo

6 subjects were given a drug (treatment group) and an additional 6 subjects were given a placebo (control group). Their reaction time to a stimulus was measured (in milliseconds). Is the mean reaction time slower (e.g., greater) in the treatment than the control group?




```
t.test(x=treat, y=cont, alternative = "g")
```

`alternative = "g"`
tests if x has a greater mean than y

the mean reaction time of the treatment is significantly slower than the control group ($t = 3.48$, df = 10, $p = 0.003$)

4 – crops

The yearly crop yield for a particular farm is typically measured as the amount of the crop produced per acre (e.g., pounds per acre). The normal distribution can characterize crop yields over time.

Historical data indicate that next summer's cotton yield for an average Georgia farmer can be characterized by a normal distribution with a mean of 1,500 pounds per acre and σ of 250. The farm will be profitable if it produces at least 1,600 pounds per acre.

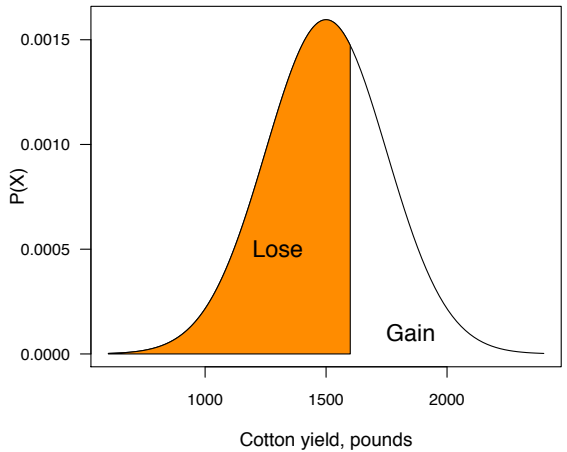
1. what is the probability that the farm will lose money next summer?
2. assume the same normal distribution is appropriate for describing cotton yield in each of the next two summers. also assume that the two yields are statistically independent. what is the probability that the farm will lose money for two straight years?

4 – crops

$N(\mu = 1500, \sigma = 250)$: profitable at 1600 pounds per acre

A. what is the probability that the farm will lose money next summer?
B. assume the same normal distribution is appropriate for describing cotton yield in each of the next two summers and that the two yields are statistically independent. What is the probability that the farm will lose money for two straight years?

```
Z <- (1600-1500)/(250)
pnorm(q=Z, mean=0, sd=1)
P(lose) = 0.655

pnorm(1600, mean = 1500, sd = 250)
P(lose) = 0.655

P(lose) x P(lose)=0.655 x 0.655 = 0.43
```



5 – fertilizer

The EPA is evaluating the claim that a new "green" fertilizer increases production of wheat. Unfertilized fields produce a mean of 40 bushels/acre. A sample of 15 1-acre fertilized fields produced a mean of 45 bushels/acre with a standard deviation of 10.

1. does the fertilizer significantly improve production?
2. what is the 95% CI of the new mean production?

5 – fertilizer

1. does the fertilizer significantly improve production?

```
t <- (45-40)/(10/sqrt(15))
[1] 1.936492

pt(t, df = 14, lower.tail = F)
[1] 0.03662922
```

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$df = n - 1$$

yes, the new fertilizer significantly increases production (t=1.94, df = 14, p=0.037)

2. what is the 95% CI of new mean production?

```
ci <- c(45 - qt(0.975, df = 14)*(10/sqrt(15)),
        45 + qt(0.975, df = 14)*(10/sqrt(15)))
ci
[1] 39.46218 50.53782
```

6 – mice

40 mice received a treatment over 3 months. We want to know whether the treatment had an impact on the weight of the mice. The weight of the 40 mice was measured before and after the treatment.

1. what type of test do you need to conduct?
2. what are your hypotheses?
3. what are the assumptions of the test?
4. what is the conclusion of your test?

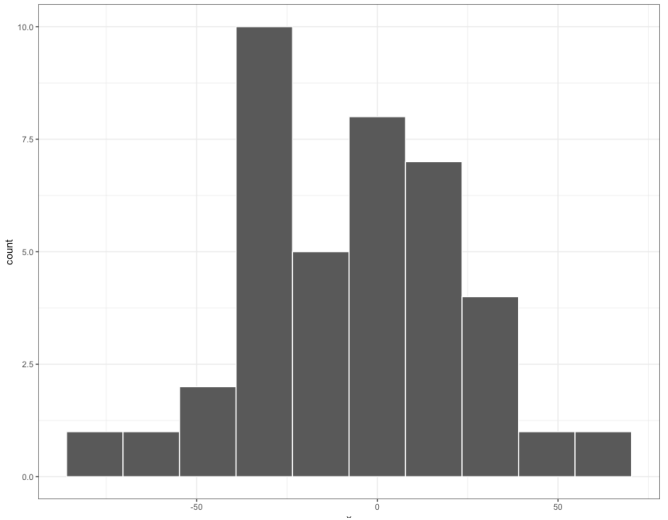
```
mice_dt <- read.csv("mice_dt.csv")
```

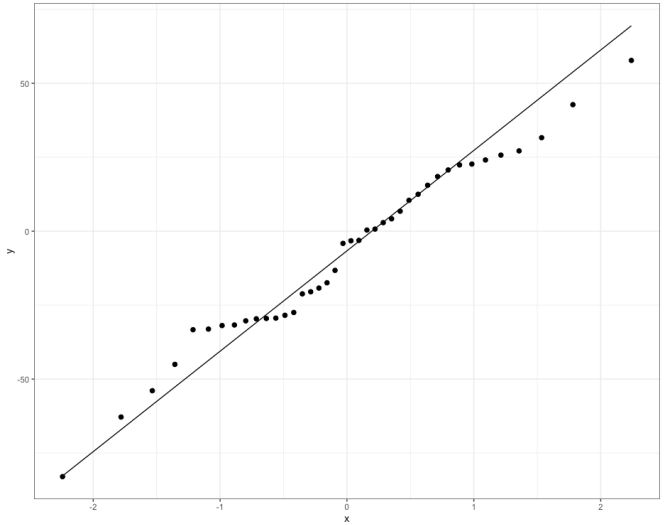
```
head(mice_dt)
  X   bf_dat   af_dat
1 1 237.3858 174.5450
2 2 207.0884 173.7672
3 3 205.4538 175.1367
4 4 209.0074 187.8218
5 5 192.1437 207.6666
6 6 167.4534 174.2017
```

6 – mice

assumptions: observations are independent, no extreme "outliers", difference between pairs are approximately normally distributed

```
mice_diff <- with(mice_dt, af_dat - bf_dat)

ggplot(data.frame(x=mice_diff), aes(x = x)) +
 geom_histogram(colour = "white", bins = 10) +
  theme_bw()

ggplot(data.frame(x=mice_diff), aes(sample=x)) +
        stat_qq() + stat_qq_line() +
        theme_bw()
```





6 – mice

40 mice received a treatment over 3 months. We want to know whether the treatment had an impact on the weight of the mice. The weight of the 40 mice was measured before and after the treatment.

```
t.test(x=before, y=after, paired=TRUE)
```

```
Paired t-test

data:  mice_dt$bf_dat and mice_dt$af_dat
t = 1.609, df = 39, p-value = 0.1157
alternative hypothesis: true mean difference
is not equal to 0
95 percent confidence interval:
 -1.960202 17.210339
sample estimates:
mean difference
       7.625068
```

no, the treatment did not significantly increase the weight of the mice (t=1.609, df=39, p=0.116)

Questions?