STAT 515 fa 2023 Lec 16 slides

Two-sample testing

Karl Gregory

University of South Carolina

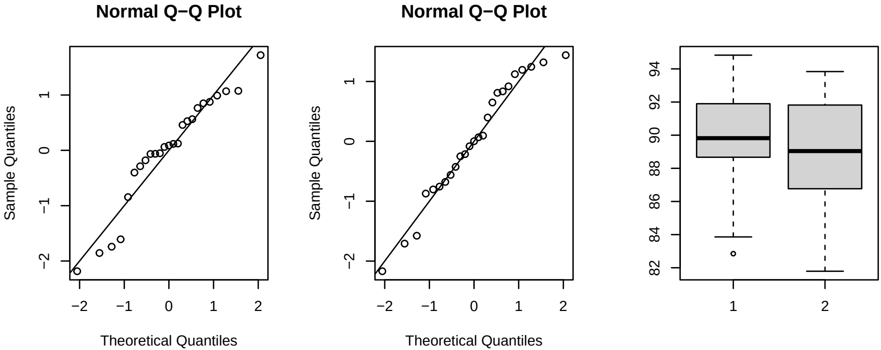
These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Think about comparing two populations:

- Compare μ_1 with μ_2 by comparing $\bar{X}_1$ and $\bar{X}_2$.
- Compare p_1 with p_2 by comparing $\hat{p}_1$ and $\hat{p}_2$.

Exercise: We wish to know whether drug tablets produced at two different sites have the same average concentration of the drug. (Ex 6.92 in [2]).

Site 1			Site 2		
91.28	86.96	90.96	89.35	87.16	93.84
92.83	88.32	92.85	86.51	91.74	91.20
89.35	91.17	89.39	89.04	86.12	93.44
91.90	83.86	89.82	91.82	92.10	86.77
82.85	89.74	89.91	93.02	83.33	83.77
94.83	92.24	92.16	88.32	87.61	93.19
89.93	92.59	88.67	88.76	88.20	81.79
89.00	84.21		89.26	92.78	
84.62	89.36		90.36	86.35	


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
82
84
86
88
90
92
94
1
2

Goals:

1. Build confidence intervals for $\mu_1 - \mu_2$.
2. Test null and alternate hypotheses about $\mu_1 - \mu_2$.

Exercise: Write down the null and alternate hypotheses for the following:

1. Do honors grads earn more in first post-grad year than non-honors grads?
2. Does a B.A versus a B.S. make a difference, on average, in salary?
3. Does a fertilizer increase crop yields?

Sampling distribution of difference in sample means

If both populations are Normal, then $\bar{X}_1 - \bar{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Take $\bar{X}_1 - \bar{X}_2$ into the "Z-world" with

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Taking $\bar{X}_1 - \bar{X}_2$ into the "t-world"

If both populations are Normal, then

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{approx}}{\sim} t_{\nu^*},$$

where

$$\nu^* = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 \left[\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}\right]^{-1}.$$

The quantity ν^* is the df of the closest "t-world". From Welch/Satterthwaite.

Taking $\bar{X}_1 - \bar{X}_2$ into the "t-world"

If both populations are Normal with equal variances $\sigma_1^2 = \sigma_2^2$, then

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where

$$S^2_{\text{pooled}} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The quantity S^2_{pooled} is a pooled estimator of the common variance when $\sigma_1^2 = \sigma_2^2$.

Confidence intervals for $\mu_1 - \mu_2$ when both populations are Normal

A $(1-\alpha) \times 100\%$ CI for $\mu_1 - \mu_2$ is given by

$$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,\alpha/2} S_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad \text{if } \sigma_1^2 = \sigma_2^2$$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu^*,\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad \text{if } \sigma_1^2 \neq \sigma_2^2.$$

It is always safe to use the second one; use the first only if $\sigma_1^2 = \sigma_2^2$ is plausible.

Exercise: For the drug concentration data, build a 95% confidence interval for the difference in means assuming first $\sigma_1^2 = \sigma_2^2$ and then $\sigma_1^2 \neq \sigma_2^2$.

```
site1 <- c(91.28,92.83,89.35,91.90,82.85,94.83,89.93,89.00,84.62,
           86.96,88.32,91.17,83.86,89.74,92.24,92.59,84.21,89.36,
           90.96,92.85,89.39,89.82,89.91,92.16,88.67)
site2 <- c(89.35,86.51,89.04,91.82,93.02,88.32,88.76,89.26,90.36,
           87.16,91.74,86.12,92.10,83.33,87.61,88.20,92.78,86.35,
           93.84,91.20,93.44,86.77,83.77,93.19,81.79)

n1 <- length(site1)
n2 <- length(site2)
xbar1 <- mean(site1)
xbar2 <- mean(site2)
s1 <- sd(site1)
s2 <- sd(site2)

sp <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2))
me <- qt(0.975,n1 + n2 - 2) * sp * sqrt(1/n1 + 1/n2)
d <- xbar1 - xbar2
lo <- d - me
up <- d + me
```

```
nu <- (s1^2/n1 + s2^2/n2)^2 / ((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
me2 <- qt(0.975,nu) * sqrt(s1^2/n1 + s2^2/n2)
lo2 <- d - me
up2 <- d + me
```

Build CIs and obtain p-values with `t.test()` in R. 😂

1. For $\sigma_1^2 = \sigma_2^2$, we can use

   ```
   t.test(x1, x2, var.equal=TRUE, mu=0, alternative="two.sided")
   ```

2. For $\sigma_1^2 \neq \sigma_2^2$, we can use

   ```
   t.test(x1, x2, var.equal=FALSE, mu=0, alternative="two.sided")
   ```

Change the `alternative` and `mu` arguments to test other sets of hypotheses.

Run `?t.test` to read the documentation.

```
> t.test(site1,site2,var.equal = TRUE)

Two Sample t-test

data:  site1 and site2
t = 0.57214, df = 48, p-value = 0.5699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.304376  2.341976
sample estimates:
mean of x mean of y
  89.5520   89.0332
```

```
> t.test(site1,site2,var.equal = FALSE)

Welch Two Sample t-test

data:  site1 and site2
t = 0.57214, df = 47.659, p-value = 0.5699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.304712  2.342312
sample estimates:
mean of x mean of y
  89.5520   89.0332
```

Must have both populations Normal *or* $n_1 \geq 30$ and $n_2 \geq 30$ to use these.

1. Inference about $\mu_1 - \mu_2$

2. Inference about $p_1 - p_2$

Exercise: Write down the null and alternate hypotheses for the following:

1. Do same number of honors and non-honors students pursue grad school?
2. Does a vaccine reduce the probability of getting an infection?
3. Do rural and urban voters differ in their preferences for a candidate?

Let $X_{k1}, \dots, X_{kn_k} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_k)$, $k = 1, 2$, and let $\hat{p}_1 = \bar{X}_1$, $\hat{p}_2 = \bar{X}_2$.

Sampling distribution of difference in sample proportions

For larger and larger n_1 and n_2, the quantity

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \text{ behaves more and more like } Z \sim \text{Normal}(0,1)$$

Rule of thumb: Need $\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15$ and $\min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15$.

Confidence interval for difference in proportions

An approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

provided $\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15$ and $\min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15$.

Exercise: It is reported that among the 319 adult first class passengers aboard the Titanic, 197 survived, while among the 627 adult third class passengers, 151 survived. The data are taken from [1].

Build a 95% confidence interval for the difference in the "true" proportions as a way of assessing whether the probability of surviving was affected by class.

Tests about $p_1 - p_2$

Define the test statistic

$$Z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Then for n_1, n_2 large, the following tests have (approx) $P(\text{Type I error}) \leq \alpha$.

H_0: $p_1 - p_2 \geq 0$ H_1: $p_1 - p_2 < 0$	H_0: $p_1 - p_2 = 0$ H_1: $p_1 - p_2 \neq 0$	H_0: $p_1 - p_2 \leq 0$ H_1: $p_1 - p_2 > 0$
Reject H_0 if $Z_{\text{test}} < -z_\alpha$	Reject H_0 if $\lvert Z_{\text{test}} \rvert > z_{\alpha/2}$	Reject H_0 if $T_{\text{test}} > z_\alpha$
p-val $= P(Z < Z_{\text{test}})$	p-val $= 2 \cdot P(Z > \lvert Z_{\text{test}} \rvert)$	p-val $= P(Z > Z_{\text{test}})$

In the above $\hat{p}_0 = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$.

Exercise: Suppose that in random samples of size 1000 of 15-17 yr-olds and 25-35 yr-olds, 6% and 3%, respectively, were found to have used JUUL in the last month. You wish to know if the proportion is higher in the younger age group. This exercise is based on some summary statistics given in [3].

1. Give the hypotheses of interest.
2. What is our conclusion at the $\alpha = 0.01$ significance level?

- Robert J MacG Dawson.
  The "unusual episode" data revisited.
  Journal of Statistics Education, 3(3), 1995.

- J.T. McClave and T.T. Sincich.
  Statistics.
  Pearson Education, 2016.

- Donna M Vallone, Morgane Bennett, Haijun Xiao, Lindsay Pitzer, and Elizabeth C Hair.
  Prevalence and correlates of juul use among a national sample of youth and young adults.
  Tobacco Control, 2018.