Lecture 7 Wednesday January 25

1) Lecture Notes pp. ~~24~~ 26 – 30

2) Hw 02 Comment:
Cite rules on p. 18 & p. 20 of the lecture notes, to justify your work.

Comparison of means from two normal samples (any sample sizes)

Basic Framework

Assume:

1. The Y's are a random sample of size n_Y from a normal population with mean μ_Y and SD σ_Y, both unknown; $\bar{Y}$ is the sample mean and S_Y is the sample SD.
2. The Z's are a random sample of size n_Z from a normal population with mean μ_Z and SD σ_Z, both unknown; $\bar{Z}$ is the sample mean and S_Z is the sample SD.
3. The two samples are independent of one another.
4. The two standard deviations are equal; that is, $\sigma_Y = \sigma_Z$.

Want:

1. A confidence interval for $\mu_Y - \mu_Z$.
2. Hypothesis tests concerning $\mu_Y - \mu_Z$.

We want a **confidence interval** (CI) for $\mu_Y - \mu_Z$. This is an interval estimate of a population parameter. Width of interval is determined by: variance of observations in both groups, sample sizes, confidence level.

True or false: *The higher the confidence level, the narrower the confidence interval.*

Also consider how variance and sample size affect width of the CI.

We also want **hypothesis tests** about $\mu_Y - \mu_Z$. We can test for any value of the parameter $\mu_Y - \mu_Z$, but we almost always test whether it is zero.

Three forms of hypothesis test about $\mu_Y - \mu_Z$:

- H_0: $\mu_Y - \mu_Z = 0$ vs. H_a: $\mu_Y - \mu_Z > 0$
- H_0: $\mu_Y - \mu_Z = 0$ vs. H_a: $\mu_Y - \mu_Z < 0$
- H_0: $\mu_Y - \mu_Z = 0$ vs. H_a: $\mu_Y - \mu_Z \neq 0$

HERE WE EXPLAIN THE THEORY THAT'S ON THE NEXT PAGE.

Goals: Construct a confidence interval and hypothesis test for $\mu_Y - \mu_Z$.

We use the statistic $\bar{Y} - \bar{Z}$ to estimate $\mu_Y - \mu_Z$.

First note that $\bar{Y} - \bar{Z}$ is a linear combination of two normal random variables:

$$\bar{Y} - \bar{Z} = (1)\bar{Y} + (-1)\bar{Z}$$

Now, find $E(\bar{Y} - \bar{Z})$ and $\mathrm{Var}(\bar{Y} - \bar{Z})$ using the rules on p. 18, and the results for the sample mean (p. 20, Lecture 5).

First, write $E(\bar{Y} - \bar{Z}) = (1)E\bar{Y} + (-1)E\bar{Z} = \mu_Y - \mu_Z$

So, $\bar{Y} - \bar{Z}$ is an unbiased estimator of $\mu_Y - \mu_Z$.

Next consider the variance:

$$\mathrm{Var}(\bar{Y} - \bar{Z}) = (1)^2 \frac{\sigma_Y^2}{n_Y} + (-1)^2 \frac{\sigma_Z^2}{n_Z} = \frac{\sigma_Y^2}{n_Y} + \frac{\sigma_Z^2}{n_Z} = \sigma^2 \left(\frac{1}{n_Y} + \frac{1}{n_Z} \right)$$

because $\bar{Y}$ and $\bar{Z}$ are independent random variables, and because ...

Finally, we need to estimate σ^2, and it turns out that the best way to do this is with the "pooled estimate."

Eg. If $n_Y = n_Z = n$, then $\hat{\sigma}^2 = \frac{1}{2}(S_Y^2 + S_Z^2)$

the average of the two sample (pooled estimate) variances.

The standard two-sample t procedures: Theory

Basic Facts

1. $E(\bar{Y} - \bar{Z}) = \mu_Y - \mu_Z$
2. $\text{SD}(\bar{Y} - \bar{Z}) = \sigma\sqrt{\dfrac{1}{n_Y} + \dfrac{1}{n_Z}}$
3. $\bar{Y} - \bar{Z}$ has a normal distribution.
4. $$\frac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\sigma\sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}} \sim \mathcal{N}(0, 1).$$
5. $$\frac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\hat{\sigma}\sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}} \sim t_{n_Y + n_Z - 2},$$

where $$\hat{\sigma}^2 = \frac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}$$

general formula for the pooled estimate of variance, whether sample sizes are equal or not

Note the general structure of the quantity in Basic Fact 5:

$$\frac{\text{Estimate} - \text{Parameter}}{\text{St. error of Estimate}}$$

This is the same structure as for the one-sample t procedures, so the CI for $\mu_Y - \mu_Z$ will have the same structure as before, also:

$100(1-\alpha)\%$

$$\text{Estimate} \pm t_{df,\,1-\frac{\alpha}{2}} (\text{SE of Estimate})$$

$$\bar{Y} - \bar{Z} \pm t_{n_Y + n_Z - 2,\,1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}$$

Let's move on to create the hypothesis test.

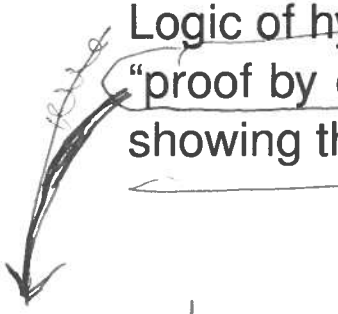
Recall, we have $H_0 : \mu_Y - \mu_Z = 0$ vs. $H_a : \mu_Y - \mu_Z \neq 0$

Goal: Choose between H_0 and H_a.

First step: Form a test statistic which measures how far the data are from the null hypothesis value of the parameter.

Numerator of T_{obs} is $\bar{Y} - \bar{Z} - (\mu_Y - \mu_Z)^0$, where $(\mu_Y - \mu_Z)^0$ is the value specified by H_0, which here is zero. Numerator, then, is just $\bar{Y} - \bar{Z}$.

Logic of hypothesis testing: Assume H_0, which we are trying to disprove. Use "proof by contradiction." We want to prove H_a by first assuming H_0 and then showing that the data contradict this assumption.



Ancient method of proof

Like a jury trial: First assume "not guilty"
Put burden of proof on prosecuting attorney to establish guilt "beyond a shadow of a doubt."

The standard two-sample t-test:

Form

$$T = \frac{\bar{Y} - \bar{Z}}{\sqrt{\left(\frac{1}{n_Y} + \frac{1}{n_Z}\right)\frac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}}}.$$

If the null hypothesis is true, and the two population variances are equal, then the distribution of this is t with $n_Y + n_Z - 2$ df.

This is called the "null distribution."

Explanation of use of P-value in "proof by contradiction"
We have a test statistic:

$$T_{\text{obs}} = \frac{\bar{Y} - \bar{Z}}{\text{SE}(\bar{Y} - \bar{Z})}$$

which has a known *null distribution*.

General def.

Null distribution: This is the sampling distribution of the test statistic if the null hypothesis is true.

The P-value is computed as "the probability of a value of the test statistic as extreme or more so than you observed, assuming the null hypothesis."

Interpretation of P-value. The P-value is a measure of how unusual our data is if H_0 is true. "Small P is bad for null."
This P-value is used to establish that we have (or don't have) a contradiction of H_0. "Small P is good for alternative."

"P-value" is not interpreted as a probability