Spring 2021

SKKU Biostats and Big data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Lecture 17
Inferences about means

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

Chapter 20: Testing Hypotheses About Proportions

- **Hypothesis testing:** the logic of jury trial
- **P-value:** the probability of seeing the observed data given that the null hypothesis is true.
- P-value is for decision-making, but it should not be blinded. It's context-dependent.
- Many issues related to P-values: *"Small P value can make a hypothesis more plausible, but the difference may not be dramatic."*
- Problems in misuses of P-value are just the tip of the iceberg.
- Let's not fool ourselves!

Central Limit Theorem for Means

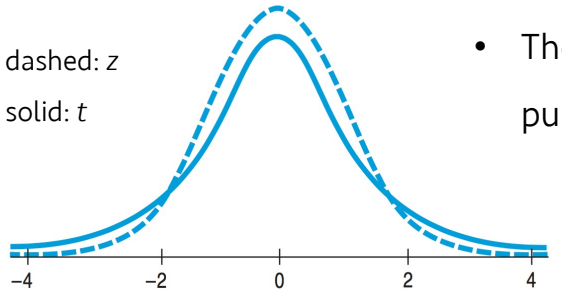
- For quantitative data, sample mean, $\bar{y}$: $Mean(\bar{y}) = \mu, SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$
- How can we know σ?
  - We use s, sample standard deviation based on data.
  - Then, standard error: $SE(\bar{y}) = \frac{s}{\sqrt{n}}$

- Normal models for SE, i.e., $\bar{y} \pm z^* SE(\bar{y})$ work for large sample sizes well.
- But people began to notice problems with smaller samples.

Gosset's t

- William S. Gosset (1876-1937)

- Chief Experimental Brewer for the Guinness Brewery in Dublin, Ireland
- His sample sizes were very small – often 3 or 4.
- He found that when he used the standard error, $\frac{s}{\sqrt{n}}$, as an estimate of the standard deviation of the mean, the shape of the sampling model changed.

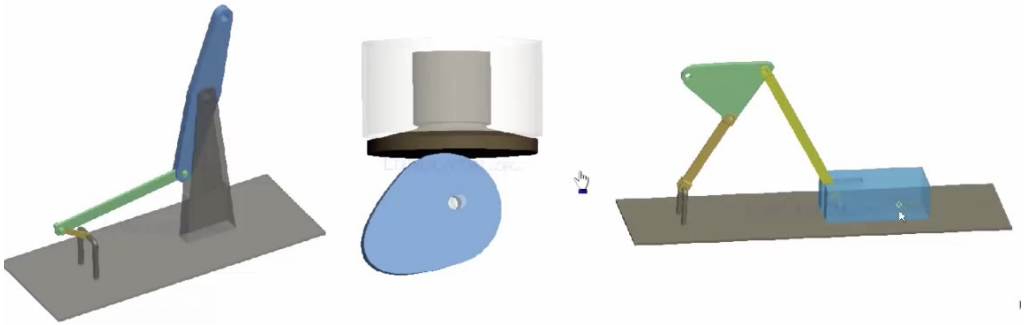


- The Guinness Company did not allow their employees to publish, so he published the paper under the pseudonym, "Student": thus, **Student's *t***

Degrees of freedom

- Gosset's sampling distribution model is always bell-shaped, but the tail was a little bit **heavier** with **small samples**. This was related to a parameter, called degrees of freedom.

- Degrees of freedom:



- The number of values that are free to vary after we've estimated parameters.
- E.g., we have n data points, and then have estimated mean, $\bar{y}$, then degrees of freedom (df) = n - 1.

Why standard deviation is divided by *n* – 1 (finally)

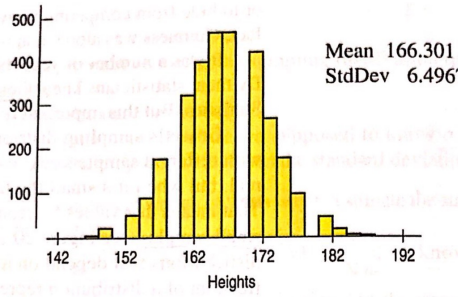
- If we know the true population mean, μ, the sample standard deviation is $s = \sqrt{\frac{\sum(y-\mu)^2}{n}}$
- Given that we don't know μ, we are using sample mean, $\bar{y}$, instead.
- Then, the data value tend to be close to their own sample mean, not the population mean (i.e., degrees of freedom decreased).
- Then, our estimated standard deviation is always (systematically) smaller than the true standard deviation.
- By dividing $\sum(y-\bar{y})^2$ by *n-1* (degrees of freedom) instead of by *n*, we can fix the bias (i.e., making the standard deviation larger).
- When sample size is *small*, the difference is much more important.

Quiz 17-01

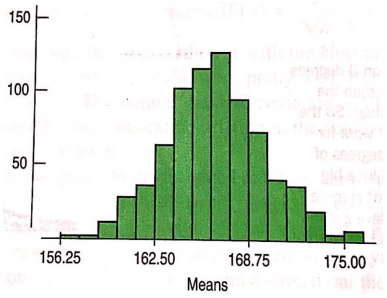
https://forms.gle/W357DqpEfRB58nro7

What Gosset see:

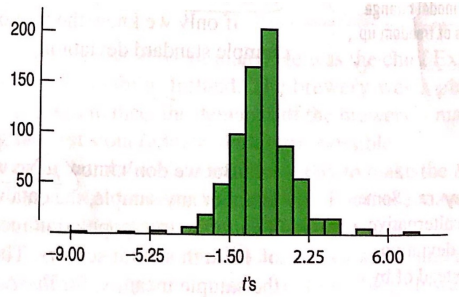
Gosset's simulation (sample n = 4)



Computer simulation

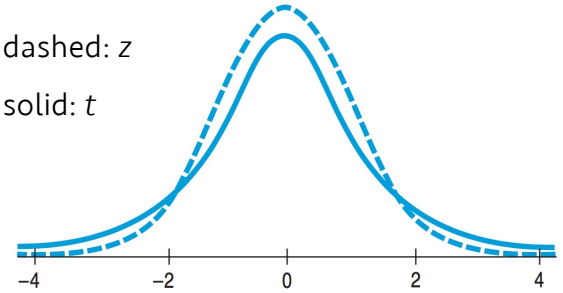


When normal model is used



dashed: z

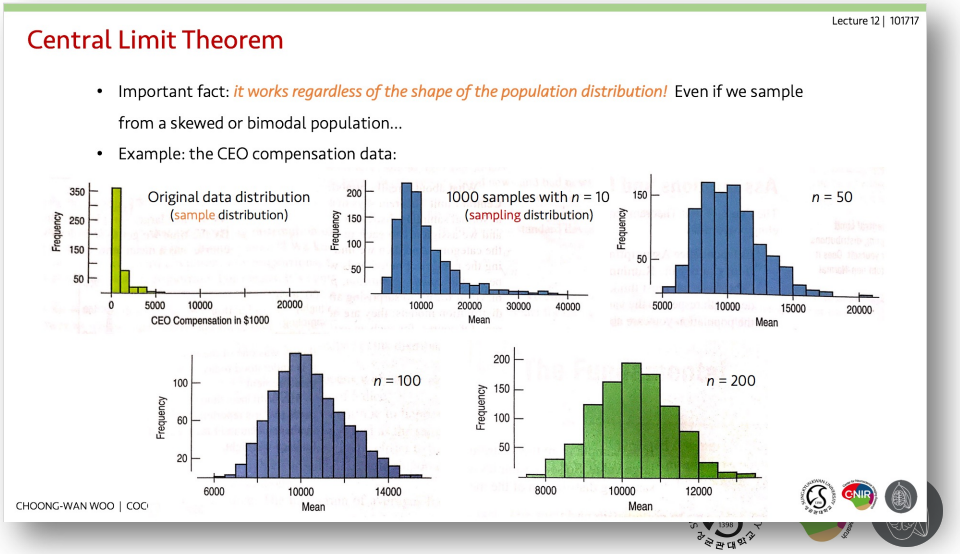
solid: t



A Confidence Interval for Means

- Confidence interval for Means: $\bar{y} \pm t^*_{df} \times SE(\bar{y})$, where $SE(\bar{y}) = \frac{s}{\sqrt{n}}$, and $df = n - 1$
  - t^* instead of z^*
  - Student's t model are unimodal, symmetric, and bell-shaped similar to the Normal.
  - As the degrees of freedom increase, the t-models look more and more like the Normal.

- **Assumptions and conditions:**
  - Independence assumption
  - **Normal population assumption:**
    - Student's t-models won't work for small samples that are badly skewed.
    - If n > 40 or 50, t methods are safe to use in most of cases.



Interpreting Confidence Intervals

College students' sleep hour: Mean = 6.64, 90% confidence interval [6.272 7.008]

- Don't say:
  - Confidence interval is about the *mean*, not *individual* students.
    - "*90% of all students* sleep between 6.272 and 7.008 hours per night".
    - "We are 90% confident that *a randomly selected student* will sleep between 6.272 and 7.008 hours per night".
  - "The mean amount students sleep is 6.64 hours 90% of the time."
    - Confidence interval doesn't assume that true mean varies.
- **Do say:** "90% of intervals that could be found in this way would cover the true value."

One sample t-test for the mean

- Null hypothesis, H0: $\mu = \mu_0$
- $t = \frac{\bar{y} - \mu_0}{SE(\bar{y})}$, where $SE(\bar{y}) = \frac{s}{\sqrt{n}}$
- When the conditions are met, this statistic follows a Student's t-model with n-1 degrees of freedom. We use that model to obtain a P-value.

Quiz 17-02

https://forms.gle/xBk8LWa5E4x8YDuQ6

Examples: Confidence interval

In 2004, a team of researchers published a study of contaminants in farmed salmon.[4] Fish from many sources were analyzed for 14 organic contaminants. The study expressed concerns about the level of contaminants found. One of those was the insecticide mirex, which has been shown to be carcinogenic and is suspected to be toxic to the liver, kidneys, and endocrine system. One farm in particular produced salmon with very high levels of mirex. After those outliers are removed, summaries for the mirex concentrations (in parts per million) in the rest of the farmed salmon are:

$$n = 150 \qquad \bar{y} = 0.0913 \text{ ppm} \qquad s = 0.0495 \text{ ppm}.$$

Examples: One-sample t-test

RECAP: Researchers tested 150 farm-raised salmon for organic contaminants. They found the mean concentration of the carcinogenic insecticide mirex to be 0.0913 parts per million, with standard deviation 0.0495 ppm. As a safety recommendation to recreational fishers, the Environmental Protection Agency's (EPA) recommended "screening value" for mirex is 0.08 ppm.

Confidence Intervals and Significance tests

- Built from the same calculations
- Complementary ways of looking at the same question from two different perspectives

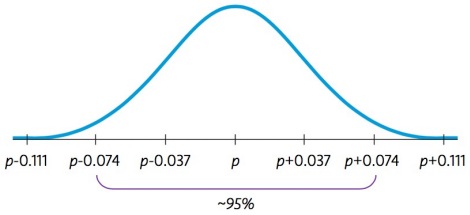
- Hypothesis tests
  - start with a proposed *parameter value* and ask if the *data* are consistent with that value
- Confidence intervals
  - start with the *data* and finds an interval of *plausible values* for where the parameter may lie
  - The confidence interval contains all the null hypothesis values we can't reject with these data.

- Both use the same standard error of the statistic as a ruler
  - but except for a *proportion*
  - $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ is different from $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$ based on the null hypothesis (p)

$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ is different from $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$ based on the null hypothesis (p)

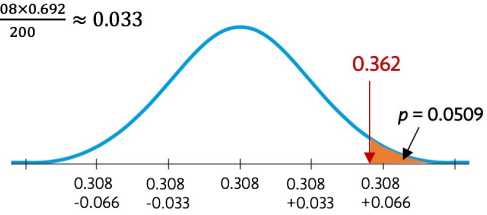
Standard error

- When we estimate the standard deviation of a sampling distribution, we call it a standard error.
- For a sample proportion, $\hat{p}$, the standard error is $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- For the Facebook users, $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.308 \times 0.692}{156}} = 0.037 = 3.7\%$



Hypotheses

- **Hypothesis**: questions like, has the Facebook users who update their status daily increased since last month?
- **Null hypothesis:** null because it assumes no changes, thus p = 30.8%
- **Alternative hypothesis**: H_A: $p > 30.8\%$
- We observed a new $\hat{p}$ from 200 respondents.
- Based on the null hypothesis, $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.308 \times 0.692}{200}} \approx 0.033$
- Let's say the observed $\hat{p}$ = 36.2%
- Then, $z = \frac{0.362 - 0.308}{0.033} = 1.6364$
- p = 0.0509 (one-tail)



Choosing the Sample size

- $ME = t^*_{n-1} \times SE(\bar{y})$, where $SE(\bar{y}) = \frac{s}{\sqrt{n}}$
- $ME = t^*_{n-1} \frac{s}{\sqrt{n}}$

- But before collecting data, we don't know *s*.
- In addition, without knowing *n*, we can't calculate *t**.

- For *s*, we can use previous studies or run a pilot study.
- For *t**, we can use *z** instead, if our estimated sample size is 60 or more.

- If the target sample size is smaller than that, you can use *z** first, and then finding *n*, then replacing *z** with *t** for *n*, and calculate *t** correctly again.

Quiz 17-03

https://forms.gle/WHUhuhoHyaBrnEjKA

Key Points

Chapter 21: Inferences About Means

- Central Limit Theorem: $Mean(\bar{y}) = \mu, SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$, standard error, $SE(\bar{y}) = \frac{s}{\sqrt{n}}$
- Degrees of freedom: the number of values that are free to vary after we estimate parameters.
- Confidence interval for Means: $\bar{y} \pm t^*_{df} \times SE(\bar{y})$, where $SE(\bar{y}) = \frac{s}{\sqrt{n}}$, and $df = n - 1$ (t^* instead of z^*)
- Hypothesis tests and confidence interval are built from the same calculations, and looking at the same question from two different perspectives.
- Hypothesis tests start with a proposed *parameter value* and ask if the *data* are consistent with that value, while confidence intervals start with the *data* and finds an interval of *plausible values* for where the parameter may lie.