


Advanced Experimental Design


Topic Four
Hypothesis testing (z and t tests) &
Power



Agenda

- Hypothesis testing
 - Sampling distributions/central limit theorem
 - z test (σ known)
 - One sample z & Confidence intervals
 - t test (σ known)
 - Degrees of freedom and the t distribution
 - One sample, Paired & Independent samples
 - Power

2



Sampling distributions

- Draw all possible random samples of size n from a population of N
- Calculated the mean of each sample
- We have created a sampling distribution of means
- The mean of this sampling distribution of means will equal the population mean

3



Central limit theorem

- o In a population with a mean μ and a standard deviation σ
- o distribution of sample means approaches a normal distribution with a mean $\mu_{\bar{x}}$ and standard deviation of $\frac{\sigma}{\sqrt{n}}$
- o If n is sufficiently large, the sampling distribution will be approximately normal regardless of the shape of the population distribution
- o What is sufficiently large?

4



Sampling distribution of mean

- o A normal distribution, characteristics all apply
- o 68% of all sample means will be within ± 1 s.d. from $\mu_{\bar{x}}$
- o 95% between ± 1.96 s.d., total of 5% below -1.96 and above 1.96
- o Can say that there is 2.5% of the area in each tail

5



Standard Error of the Mean

- o Standard error of the mean = s.d. of the sampling distribution of the mean
- o $\sigma_{\bar{x}}$ (or SEM) = $\frac{\sigma}{\sqrt{n}}$
- o Lets us calculate probability of sample means
- o What happens to the SEM as N increases?
- o Study: Data on hours worked per week at part time jobs by 100 college students

6

● ● ● | Z-test / Sample mean probabilities

- $\bar{x} = 11.7$ and we know that the population s.d. $\sigma = 4.0$

$$\sigma_{\bar{x}} = \frac{4}{\sqrt{100}} = .4$$

- With a sample mean of 11.7 and SEM = .4 "is it possible" for the population mean, μ , to be 12?

7

● ● ● | Z test (cont.)

- New z score to answer this question $Z = \frac{x - \bar{x}}{s}$
- x = observation
- \bar{x} = sample mean
- s = sample standard deviation

$$Z = \frac{x - \bar{x}}{s}$$

- Replace variables:
- X becomes the sample mean
- \bar{x} becomes the mean of the sampling distribution
- s replaced by s.d. of the sampling dist. of mean (SEM)

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

8

● ● ● | Z test (cont.)

- $(11.7-12)/0.4 = -.3/.4 = -0.75$
- By sampling error alone, sample means above and below the mean are equally likely
- interested in the probability that a mean could vary this much or more in either direction
- What is the probability?
- What would we say about the p(sample mean of 11.7|true pop. mean is 12)?

9

Z test (cont.)

- Could μ be 12.3?
- How do we calculate this probability?
- What do we think about possibility μ is 12.3?
- Could μ be 12.5?
- What do we think about the possibility that μ is 12.5?
- Can view graphically.

10

Population parameters

- Further sample mean deviates from hypothesized pop. mean, smaller the probability the sample mean came from a population with the hypothesized mean.
- Hypothesized population means and associated probabilities of a greater deviation than our sample mean of 11.7 hours

Hypothesized Population Mean	Amount of Deviation from μ	Probability of Deviation
12 hours	0.3 hours or more	$p = .4532$
12.3 hours	0.6 hours or more	$p = .1336$
12.5 hours	0.8 hours or more	$p = .0456$

11

Confidence Intervals around the Mean

- calculate confidence intervals around our sample mean for generalizing back to the larger population
- $\bar{X} \pm 1.96(\sigma_{\bar{x}}) = 95\%$ confidence interval
- $\bar{X} \pm 2.576(\sigma_{\bar{x}}) = 99\%$ confidence interval
- $\bar{X} \pm 1.645(\sigma_{\bar{x}}) = 90\%$ confidence interval

12

● ● ● | SAT problem

- 1979 N. Dakota, 238 students took SAT verbal test, with mean of 525. No sd was reported
- Is this consistent with the idea that the SAT has a μ of 500 & $\sigma = 100$?
- Using confidence interval or z test approach will yield same answer
- Would H_0 be rejected if we were looking for evidence that the SAT scores in general had been declining?
- How about 2345 Arizona students with a mean of 524?

13

● ● ● | σ known vs. unknown

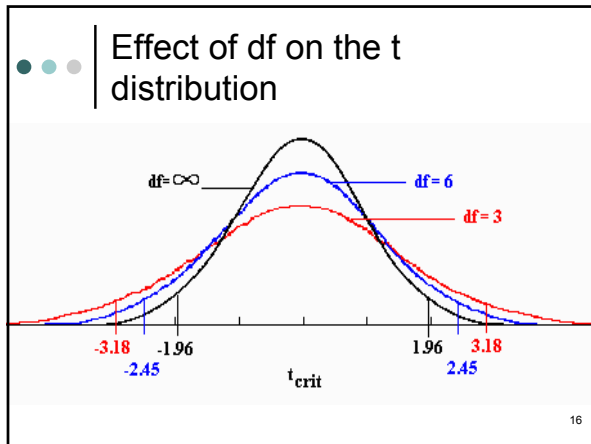
- When σ is known, we can use normal curve areas to determine the limits of our confidence intervals
- What do we do when σ is not known?
- How do we address our uncertainty about σ ?
- The t Distribution: Defined
 - The t distribution is a theoretical probability distribution
 - Symmetrical, bell-shaped, and similar to the standard normal curve
 - Differs from the standard normal curve, has an additional parameter, called degrees of freedom, which changes its shape.

14

● ● ● | Degrees of Freedom

- Degrees of freedom, symbolized by df, is a parameter of the t distribution which can be any real number greater than zero (0.0)
- Value of df defines a particular member of the family of t distributions
- A t distribution with a smaller df has more area in the tails of the distribution than one with a larger df
- Gets at the number of dimensions along which the data is free to vary

15



- ### d.f. for Different t-tests
- One Sample: $df = n - 1$
 - Paired (correlated): $df = n(\text{pairs}) - 1$
 - Ind. Samples: $df = n(\text{group 1}) + n(\text{group 2}) - 2$
 - Relationship to the Normal Curve
 - As df increases, t approaches the standard normal distribution ($\mu = 0.0$, $\sigma = 1.0$)
 - The standard normal curve is a special case of the t distribution when $df = \infty$.
 - The t distribution approaches the standard normal distribution relatively quickly, when $df = 30$ the two are almost identical.
- 17

- ### Selecting alpha (α)
- Setting the value of α is not automatic
 - depends on the relative costs of the two types of errors
 - probabilities of the two types of errors (I and II) are inversely related
 - if cost of Type I error high relative to cost of Type II, α should be set low
 - if cost of a Type I error low relative to cost of Type II, α should be higher
- 18

One sample t-test

- Used when we want to compare a mean from a single sample against a hypothesized population mean.
- Basically creating a ratio with the deviation between the observed mean and hypothesized mean in the numerator and the standard error of the sample mean in the denominator.
- We compare this statistic with the critical value of t for the appropriate df and our chosen α level.
- If the statistic exceeds critical value, reject null.

19

Independent and Paired (correlated) samples t tests

- basic idea is taking a difference observed, divide by standard error of the difference giving a t statistic
- Three types of paired (correlated) designs:
 - Natural pairs
 - Matched pairs
 - Repeated measures
- Independent samples
 - no way to pair observations

20

SPSS set up differences

- One sample t, obviously only one variable is used
- Matched samples design, data must be entered in SPSS preserving the pairing that is present in the sample in two variables, each containing one group (or time) of the dependent variable
- Independent samples, all of the DV data goes in one column and another column is used for entering the grouping variable

21

Hypothesis testing steps

- State your research hypothesis
- State your null and alternative hypotheses
- Select your acceptable level of significance (α level), generally 0.05
- Collect and summarize data from a sample
- Calculate the probability of the sample data if the null hypothesis is true
- Decide whether to reject or retain the null hypothesis
- Describe your results in terms of the constructs you set out to investigate

22

Comparison of One and Two-tailed t-tests

- With 10 df and $\alpha=.05$, 2 tailed $t_{crit}=2.228$, 1 tailed=1.812
 - 1. If $t_{obs} = 3.37$, then significance would be found in the two-tailed and the positive one-tailed t-tests.
 - The one-tailed t-test in the negative direction would be n.s., because α was placed in the wrong tail.
 - 2. If $t_{obs} = -1.92$, then significance would only be found in the negative one-tailed t-test
 - If the correct direction is selected, more likely to reject the null hypothesis, the significance test is said to have greater power in this case.

23

One and Two-tailed t-tests (cont.)

- Again: Selection of one or two-tailed t-test must be before the experiment is performed!!
- Not appropriate to find that $t_{obs} = -1.92$, then say "I meant to do a one-tailed t-test."
- Reviewers of articles are sometimes suspicious about one-tailed t-tests
- If there is any doubt, a two-tailed test should be done.

24

Assumptions of the t-test

- Normal distribution of dependent variable
 - Transformation can correct
 - Nonparametric test – if extreme violation
 - Not a major issue with large n and = group n
- Homogeneity (equality) of variances across groups
 - SPSS provides adjusted statistics in output

25

Levine's test

- With independent samples t test, SPSS provides Levene's test for equality of variances
- If we accept the null hypothesis that the variances are equal, we use the "equal" row of t statistic, df, etc.
- If the null hypothesis of equal variances is rejected, i.e., the p value is < 0.05 , we use the unequal row.

26

"Significance" of significant differences

- When we reject H_0 we conclude that there is a statistically significant difference between means
 - Statistically significant does not necessarily mean clinically meaningful
 - We can get highly significant results (statistically) that are completely meaningless in any practical sense
 - How might this happen?

27

● ● ● | Sample size and mean values to achieve statistical significance

Sample Size	Reader Mean	Population Mean	<i>p</i>
4	110.0	100.0	0.05
25	104.0	100.0	0.05
64	102.5	100.0	0.05
400	101.0	100.0	0.05
2,500	100.4	100.0	0.05
10,000	100.2	100.0	0.05

28

● ● ● | Power

- What is power?
- Cohen 1960 JAP power, medium eff = 0.48
- Statistical Power Analysis for the Social & Behavioral Sciences
- no complex mathematics
- Did this lead to things improving in the literature?

29

● ● ● | Power in the literature

- Sedlmeier & Gigerenzer (1989) replicated Cohen's survey
- Of 54 articles, 2 mentioned power
- None estimated power, necessary n, or anticipated effect size
- Average power did change
- 11% of studies framed research hypotheses as the null hypothesis: What is the problem with this?

30



Power practicalities

- What is minimum n/group for experiments?
- Confusion with central limit theorem – small sample statistics
- Power with “recommended” n , $\alpha = .05$, $d = .5$: 0.47
- Much better for $d = .8$, power = .8
- How big are most effects in social / behavioral sciences?

31



Power practicalities

- What do we do about this?
- Easiest: use g-power to estimate required sample size a-priori or use Howell's approximations
- We'll walk through these after going through computational details of t's.

32
