

PSYCH 710

Review of Statistical Inference

1

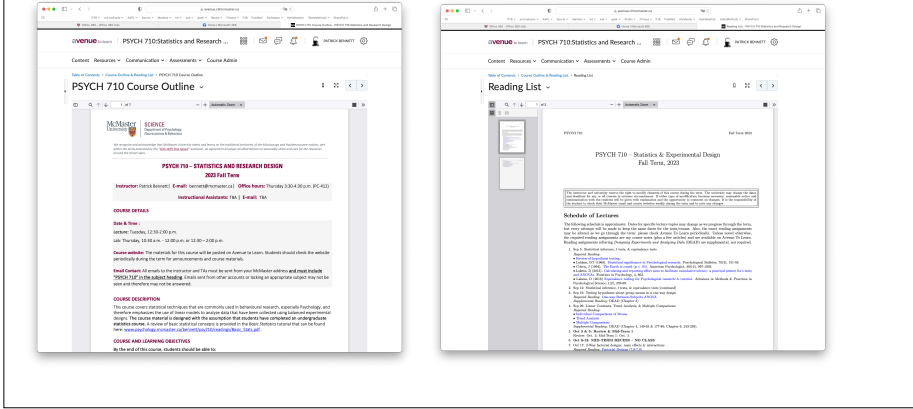
Overview

- Course Information
- What is statistics?
- Ways of collecting data
- Modes of statistical analysis
- Sampling distributions & parameter estimation
- z tests & t tests

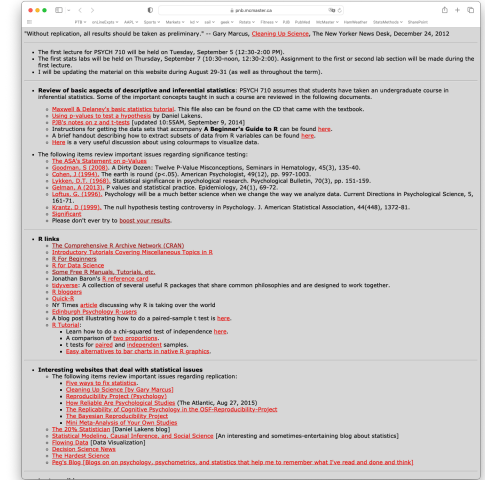
Course Management

- All lecture materials will be available on Avenue 2 Learn
 - reading assignments
 - lecture slides
 - labs & homework assignments
- Students are expected to install and use R
 - labs & exams will be in Psychology computer cluster (PC-154)
 - you may use your own laptop
- Labs, homework, & exams are open-book
 - you may use any/all aids to complete exams
 - you may collaborate on labs and homework assignments

Avenue To Learn



Avenue To Learn



R home <https://www.r-project.org>



[home]

Download

CRAN

R Project

About R

Logo

Contributors

Reporting Bugs

Conferences

Search

Get Involved: Mailing Lists

Developer Pages

R Blog

R Foundation

Board

Members

Donors

Donate

Help With R

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred **CRAN mirror**.

If you have questions about R like how to download and install the software, or what the license terms are, please read our **answers to frequently asked questions** before you send an email.

News

- **R version 4.0.2 (Talking Off Again)** has been released on 2020-06-22.
- **useR!** 2020 in Saint Louis has been cancelled. The European hub planned in Munich will not be an in-person conference. Both organizing committees are working on the best course of action.
- **R version 3.6.3 (Holding the Windsock)** has been released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a supporting member

News via Twitter

The R Foundation Retweeted

userR2020mc
@userR2020mc

Please let us know how you liked **useR2020**

Even if you joined only 1 event or watched 1 talk, you'r

CRAN

<https://cran.r-project.org>



CRAN
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages. **Windows** and **Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for Mac OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2020-06-22, Taking Off Again) [R-4.0.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha](#) and [beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features](#) and [bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

Video Tutorials on Using R

www.youtube.com/playlist?list=PLqz0L9-eJTNAARFXgwbqGo56NtbJnB37A

The screenshot shows a YouTube playlist with 17 videos. The first six videos are listed:

1. What is RStudio and Why Should You Download It? | R Tutorial 1.1 | MarinStatsLectures
2. Download and Install R and RStudio | R Tutorial 1.2 | MarinStatsLectures
3. Getting started with R: Basic Arithmetic and Coding in R | R Tutorial 1.3 | MarinStatsLectures
4. Create and Work with Vectors and Matrices in R | R Tutorial 1.4 | MarinStatsLectures
5. Import Data, Copy Data from Excel to R CSV & TXT Files | R Tutorial 1.5 | MarinStatsLectures
6. Importing/Reading Excel data into R using RStudio (readxl) | R Tutorial 1.5b | MarinStatsLectures

Video Tutorials on Using R

www.youtube.com/playlist?list=PLqz0L9-eJTNAARFXgwbqGo56NtbJnB37A

The screenshot shows a YouTube video player for the video 'Writing Scripts of Code in R'. The video title is 'Writing Scripts of Code in R' and the description includes 'and in this video we'll talk about writing scripts of code'. The video has 165,517 views and was uploaded on August 9, 2013.

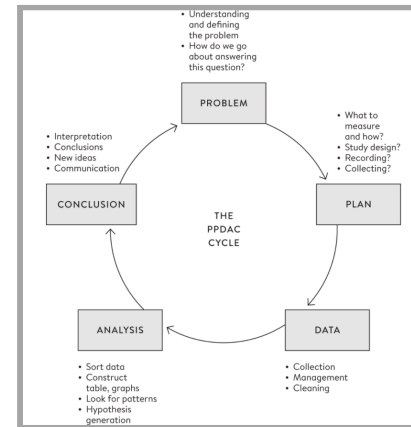
Review of Statistical Inference

What is statistics?

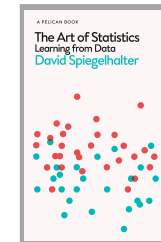
Role of statistics in research

- Statistical methods help us to collect, organize, summarize, analyze, interpret, & present data

Role of statistics in research (PPDAC)



David Spiegelhalter



14

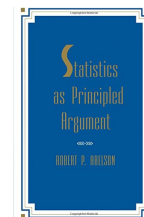
Role of statistics in research

- Statistical methods help us to collect, organize, summarize, analyze, interpret, & present data
- “The role of statistics is not to discover truth. The role of statistics is to resolve disagreements between people.” - Milton Friedman

Role of statistics in research

“...the purpose of statistics is to organize a useful argument from quantitative evidence, using a form of principled rhetoric*.” - Robert P. Abelson

*rhetoric: the art of effective/persuasive speaking or writing



16

Ways of collecting data

Ways of Collecting Data

- Designed Experiments -
 - effects of independent variables on dependent variables
 - random assignment of “subjects” to conditions
- Correlational Studies -
 - associations among predictor & criterion variables
 - “subjects” come with their own set of variables
- Both types can be combined into a single study/analysis (e.g., ANCOVA)

18

Correlational Studies

- Measure the association between predictor & criterion variables
- Predictor variables are not manipulated by investigator
 - each event/subject comes with own set of variables
 - but values on variables differ across events/subjects
- Difficult to establish causal relation between variables

19

Designed Experiments

- Measure causal effects of independent variables on dependent variables
- Independent variables usually manipulated by experimenter
 - not always (e.g., “age” in developmental studies)
- Whenever possible, participants should be assigned randomly to conditions

20

Random Assignment

- In Psychology, designed experiments use subjects that also come with their own set of intrinsic characteristics
- These characteristics (personality, motivation, intelligence, etc.) probably affect dependent variable
- HOWEVER, subjects in most designed experiments are randomly assigned to experimental conditions
- So, effects of subject differences should be UNRELATED TO EFFECTS OF INDEPENDENT VARIABLES
 - big advantage of designed experiments over correlational studies

21

Modes of Statistical Analysis

Descriptive vs. Inferential
Exploratory vs. Confirmatory

22

Descriptive vs. Inferential Statistics

- Descriptive statistics:
 - describes important characteristics of the sample
 - uses graphs & statistics e.g., mean or standard deviation
- Inferential statistics:
 - uses sample to make claims about a population
 - e.g., estimate population parameters from sample statistics
 - e.g., investigate differences among population by examining differences among samples [effect size & association strength]

23

Exploratory vs Confirmatory Analyses

- Exploratory Data Analysis
 - first major proponent was John Tukey
 - goal: discover & summarize interesting aspects of data
 - discover interesting hypotheses to test
- Confirmatory Data Analysis
 - data are gathered & analyzed to evaluate specific a priori hypotheses
 - example: clinical drug trials
- Important not to confuse two types of analyses
 - replication crisis in Psychology related to confusion about two types of research



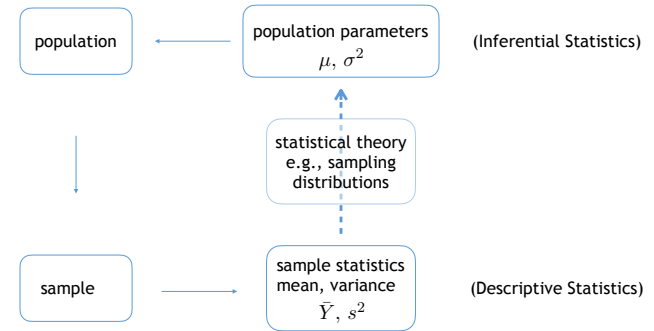
John Tukey

24

Sampling Distributions & Point Estimators

25

Descriptive & Inferential Statistics



26

Inference: Samples to Populations

- Population: all events (subjects, scores, etc)) of interest
- Sample: subset of population
 - random sample: each member of population has equal chance of being selected
 - convenience sample (e.g., psychology undergraduates)
- Inference depends on quality of relation between sample & population.
 - e.g., Is sample representative of population?

27

Sample Statistics vs. Population Parameters

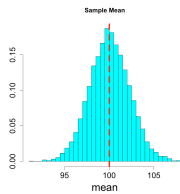
- We **can** use sample statistics to estimate population parameters
 - The sample mean, \bar{Y} , is an unbiased estimate of population mean, μ
 - The sample variance, s^2 , is an unbiased estimate of population variance, σ^2
 - [N.B. True when using (n-1) in the denominator]
 - sample standard deviation, s , is a biased estimate of population variance, σ [slightly too small]
 - The sample correlation, r , is a biased estimate of the population correlation, though the bias may be small when n is large
- What is an “unbiased” estimate?
 - If the average value of many sample statistics (e.g., \bar{Y}) equals the population parameter (e.g., μ), the statistic is an unbiased estimate of the parameter

28

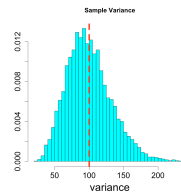
Sampling Distributions of Mean and SD

distributions of statistics
for 5000 samples (n=20)

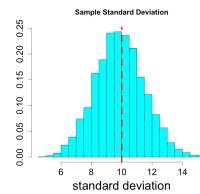
population mean = 100
population variance = 100
population sd = 10



mean = 100 (100)
variance = 5.07 (5)



mean = 100 (100)
variance = 1066 (1053)



mean = 9.87
variance = 2.65

29

Interval Estimators & Confidence Intervals

30

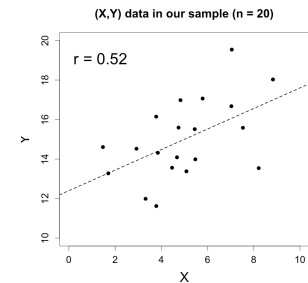
Confidence Interval

- 95% Confidence Interval
 - an interval estimate of the value of a population parameter (e.g., population correlation)
 - e.g., the population correlation lies between “r-low” and “r-high”
 - $CI_{95\%}$ is calculated from data in your sample
 - the interval varies across samples
 - we wouldn't expect it to be exactly the same for each random sample of (X,Y) values
 - in the long run, the interval contains the true population value 95% of the time
- how can we calculate $CI_{95\%}$ for our correlation, r?
 - there are several methods... we first demonstrate the percentile bootstrap method
 - N.B. the method is not as important as understanding the meaning of the $CI_{95\%}$

31

calculating 95% confidence interval for ρ

- our sample $r = 0.52$
- what is $CI_{95\%}$ for ρ ?
- percentile bootstrap method:
 - uses (X,Y) sample as estimate of (X,Y) population
 - calculate r^* on bootstrapped samples:
 - randomly select 20 (X,Y) pairs **from the data**
 - calculate r for each bootstrapped sample (r^*)
 - repeat MANY times
 - display histogram of r^*
- identify range of values containing 95% of r^*
 - range is PERCENTILE BOOTSTRAP estimate of $CI_{95\%}$ for ρ

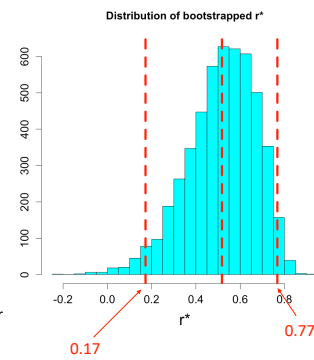


32

distribution of bootstrapped r^* values

calculate r^* for many bootstrapped samples of data ($n=20$)

- (X,Y) data is estimate of (X,Y) population
- create bootstrapped sample:
 - randomly select 20 (X,Y) pairs from data
 - calculate r for each bootstrapped sample (r^*)
 - repeat MANY times
 - display histogram of r^*
- identify range that contains 95% of r^*
 - range is PERCENTILE BOOTSTRAP estimate of 95% confidence interval for r
 - $CI_{95\%} = [0.17, 0.77]$
 - $CI_{95\%}$ is our interval estimate of population parameter ρ

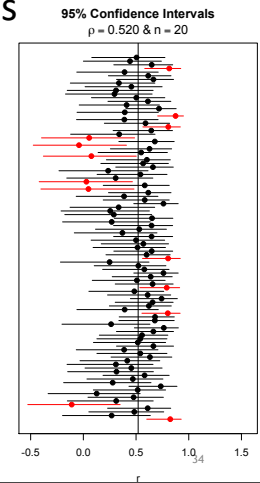


33

distribution of bootstrapped r^* values

calculate r^* for many bootstrapped samples of data ($n=20$)

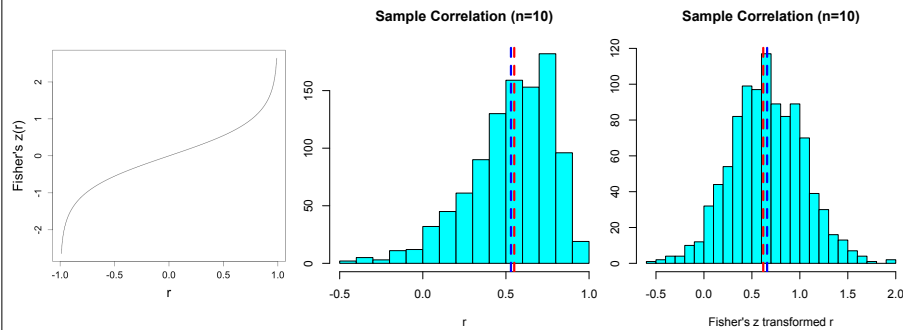
- (X,Y) data is estimate of (X,Y) population
- create bootstrapped sample:
 - randomly select 20 (X,Y) pairs from data
 - calculate r for each bootstrapped sample (r^*)
 - repeat MANY times
 - display histogram of r^*
- identify range that contains 95% of r^*
 - range is PERCENTILE BOOTSTRAP estimate of 95% confidence interval for r
 - $CI_{95\%} = [0.17, 0.77]$
 - $CI_{95\%}$ is our interval estimate of population parameter ρ



34

Confidence Interval of r

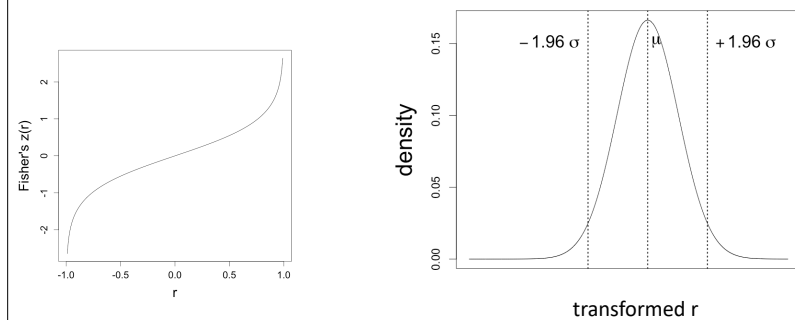
- Fisher's z transformation of r
- transformed r 's are approximately normal



35

Confidence Interval of r

- Fisher's z transformation of r
- transformed r 's are approximately normal
- calculate 95% CI by calculating critical values that capture middle 95%



36

Confidence Interval of r

1. convert r to z
2. calculate standard deviation of sampling distribution $1/\sqrt{n-3}$
3. calculate values of z that cutoff lower/upper 2.5% of distribution
4. calculate Confidence Interval defined by Fisher Z values
5. transform Z values to r values

```
1 ( sampZ <- 0.5 * log((1+ourSampR)/(1-ourSampR)) )
[1] 0.5763398
2 ( zSE <- 1/sqrt(n-3) )
[1] 0.243
3 ( zCrit <- qnorm(0.975,0,1) ) # ± zCrit
[1] 1.96
4 ( zCI <- c(sampZ-zCrit*zSE, sampZ+zCrit*zSE) )
[1] 0.1009787 1.0517008
5 ( rCI <- (exp(2*zCI)-1)/(exp(2*zCI)+1) )
[1] 0.1006368 0.7824667
```

<https://shandou.medium.com/how-to-compute-confidence-interval-for-pearsons-r-a-brief-guide-95144589c32d#>

1. Convert r to z' using Fisher's z' transform:

$$z' = 0.5 \ln \frac{1+r}{1-r}$$

2. Compute confidence intervals using the resulting z' value:

$$CI = (z' - z'_{critical}SE, z' + z'_{critical}SE)$$

where z'-critical can easily be obtained from the z-table for a given significant level, and SE is the standard error:

$$SE = \frac{1}{\sqrt{n-3}}$$

3. Convert the confident intervals in terms of z' back into r values:

$$r = \frac{e^{2z'} - 1}{e^{2z'} + 1} = \tanh(z')$$

37

Confidence Interval of r

```
1 ( sampZ <- 0.5 * log((1+ourSampR)/(1-ourSampR)) )
[1] 0.5763398
2 ( zSE <- 1/sqrt(n-3) )
[1] 0.243
3 ( zCrit <- qnorm(0.975,0,1) ) # ± zCrit
[1] 1.96
4 ( zCI <- c(sampZ-zCrit*zSE, sampZ+zCrit*zSE) )
[1] 0.1009787 1.0517008
5 ( rCI <- (exp(2*zCI)-1)/(exp(2*zCI)+1) )
[1] 0.1006368 0.7824667
```

```
> cor.test(X,Y)
```

Pearson's product-moment correlation

```
data: X and Y
t = 2.5828, df = 18, p-value = 0.01876
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1006368 0.7824667
sample estimates:
cor
0.52
```

38

z test of single observation

39

z test

A woman in the US has just given birth to a full-term baby weighing 291 kg. Is this weight unusually low?

40

Density Functions & Probability

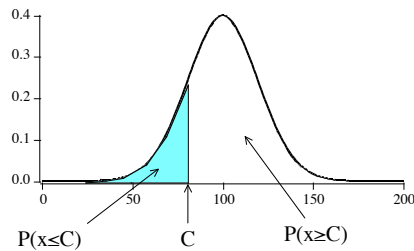


Figure 2: The probability of randomly selecting a value of x that is $\leq C$ - i.e., $P(x \leq C)$ - corresponds to the area under the probability density function that is to the left of C . $P(x \geq C)$ equals the area under the curve that is to the right of C .

41

z test

$$z = \frac{(Y - \mu)}{\sigma}$$

- z is a standardized score: # standard deviations from mean
- used to evaluate individual scores and group mean when population variance is known
- when scores or means are distributed normally
 - z is distributed normally with mean=0 and std dev=1
 - 95% of values fall between ± 1.96
 - 99% of values fall between ± 2.56

42

z test

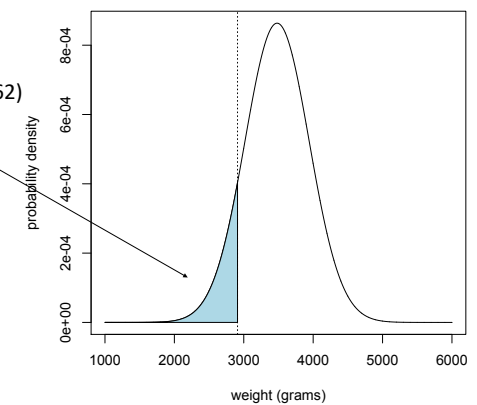
- In US, population of Caucasian (non-Hispanic) full-term infants has a mean weight of 3480 g and a standard deviation of 462 g
- The weights are distributed approximately normally
- A weight of 2910 g is 1.23 standard deviations below the mean:
 - $z = (2910 - 3480) / 462 = -1.23$
- What is the probability of observing a weight that is at least this low?

43

z test

$$\begin{aligned} p(\text{weight} < 2910) \\ &= \text{pnorm}(2910, \text{mean}=3480, \text{sd}=462) \\ &= 0.109 \end{aligned}$$

$$\begin{aligned} p(z < -1.23) \\ &= \text{pnorm}(-1.23, \text{mean}=0, \text{sd}=1) \\ &= 0.109 \end{aligned}$$



44

z test

- In US, population of Caucasian (non-Hispanic) full-term infants has a mean weight of 3480 g and a standard deviation of 462 g.
- The weights are distributed approximately normally.
- A weight of 2910 g is 1.23 standard deviations below the mean:
 - $z = (2910 - 3480) / 462 = -1.23$
- What is the probability of observing a weight at least this low?
 - $p(z < -1.23) = 0.109$

45

z test for a group mean

46

z test for means

$$z = \frac{(\bar{Y} - \mu_{\bar{Y}})}{\sigma_{\bar{Y}}}$$

- consider situation when we want to evaluate a group mean
 - e.g., measure birth weight of 100 Native-American full-term babies
 - mean = 3350 g; standard deviation = 425 g
- is group mean of 3350 g unusually low?

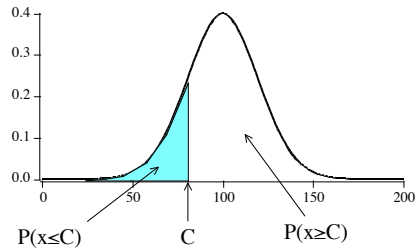
47

z test for means

- population mean & variance are known
 - mean = 3350 g; standard deviation = 425 g
- use z test
- convert observed mean to a z score: $z = \frac{(\bar{Y} - \mu_{\bar{Y}})}{\sigma_{\bar{Y}}}$

48

Density Functions & Probability



49

z test for means

$$z = \frac{(\bar{Y} - \mu_{\bar{Y}})}{\sigma_{\bar{Y}}} = \frac{(\bar{Y} - \mu)}{\sigma/\sqrt{n}}$$

- is group mean of 3350 g unusually low?
 - $n = 100$; mean = 3350 g; standard deviation = 425 g
- our null hypothesis (H_0) is:
 - sample is drawn from population with parameters that are identical for Caucasian birth weights
 - distribution of BIRTHWEIGHTS: mean = $\mu = 3480$; sd = $\sigma = 462$; distribution = NORMAL
 - distribution of MEANS: mean = 3480; sd = σ/\sqrt{n} ; distribution = NORMAL
- when H_0 is true, what is probability of getting sample mean ($n=100$) < 3350 g?
 - standard deviation of mean = Standard Error of Mean (SEM) = $462/\sqrt{100} = 46.2$
 - $z = (3350-3480) / 46.2 = -2.81$ [our mean is 2.81 standard deviations below μ]
 - $p(z < -2.81) = \text{pnorm}(-2.81,0,1) = 0.0025$
- if sample was drawn from population of Caucasian birth weights, then group mean is unusually low

50

t test for a group mean

51

why use t instead of z?

Effect of using estimate of σ

- z is defined with KNOWN population μ and σ
- only source of variation in z is sampling error of mean
- using estimate of σ introduces another source of variation in z
 - estimated z depends on sample mean AND standard deviation
- does this affect our z test?

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

$$\hat{z} = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

Effect of using estimate of population variation

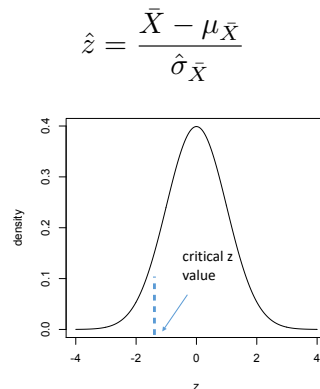


William Sealy Gosset
(aka Student)

- William Gossett applied statistics to his work in the Guinness brewery
- Under the pseudonym, Student, he investigated effects of estimating σ on z test
 - sample variance is unbiased estimate of population variance
 - but sample standard deviation is a **biased** estimate of population standard deviation
 - sample SD underestimates population SD particularly for small sample sizes
- Discovered that **using estimates of σ lead to extreme values of z more frequently than predicted** by statistical theory

Effect of inflating z score

- calculating z with estimated σ inflates z scores
- extreme z values **occur more frequently** than expected when H_0 is True
- **what effect does this have on our evaluation of H_0 ?**
 - produces more Type I errors than expected



Critical z = -1.645
 $p(z \leq -1.645 \mid H_0) = 0.05$

Effect of using estimate of population variation

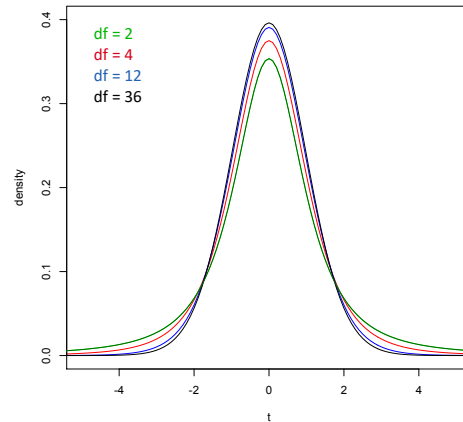


William Sealy Gosset
(aka Student)

- William Gossett applied statistics to his work in the Guinness brewery
- Under the pseudonym, Student, he investigated effects of estimating σ on z test
- Discovered that using estimates of σ lead to more “extreme” values of z than predicted by statistical theory
- Caused an increase in Type I errors
 - especially for small samples
- Devised a new test that corrected these errors
 - Student’s t test

t distribution

- unimodal
- symmetrical around zero
- has 1 parameter:
 - degrees of freedom (df)
- df alters kurtosis
 - low df associated with narrower middle portion & heavier tails
- **t approximately normal for df ≥ 35**



Back to hypothesis testing

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- When σ is NOT known
 - estimated “z” is inflated
 - our standardized score does not follow z distribution
 - using “z” increases Type I error rate
- However, standardized score DOES follow a t distribution
- Therefore, our estimated “z” actually is a t statistic
- so we use critical values of t, not z, to evaluate null hypothesis

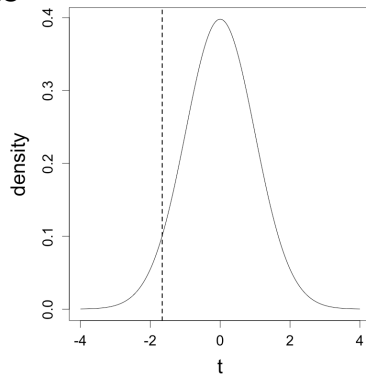
t test of sample mean

t test for means

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- consider situation when we want to evaluate a group mean
 - e.g., measure birth weight of 100 Native-American full-term babies
 - our sample:
 - mean = 3350 g; standard deviation = 425 g
- assuming our sample is drawn from typical population with UNKNOWN sigma
 - $\mu = 3480$; $\sigma = ???$; distribution's shape = ??? [we will assume it is NORMAL]
- is our sample mean of 3350 g unusually low?

t test for means



When H_0 is true, t will follow t distribution with $100-1=99$ degrees of freedom.
This t distribution is very similar to a standard normal distribution.
We expect to get a sample t below the $t=-1.66$ approximately 5% of the time.

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

t test for means

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- population mean is known (3480 g) but variance is unknown
- Sample: mean = 3350 g; standard deviation = 425 g; $n = 100$
- H_0 : our sample was drawn from typical population
 - assuming H_0 is true, is our sample mean unusually low?
- convert observed mean to a t score:
- compare to critical value of t ($t_{\text{critical}} = -1.66$)
- observed t is more extreme than t_{critical}
 - assuming H_0 is true, our mean is unusually low
 - reject null hypothesis ($p < .05$)

$$t = \frac{(\bar{Y} - \mu_{\bar{Y}})}{\hat{\sigma}_{\bar{Y}}} = \frac{3380 - 3480}{\frac{425}{\sqrt{100}}} = -2.353$$

> qt(p=.05,df=100-1)
[1] -1.660391

62

t test for means

$$t = \frac{(\bar{Y} - \mu_{\bar{Y}})}{\hat{\sigma}_{\bar{Y}}} = \frac{3380 - 3480}{\frac{425}{\sqrt{100}}} = -2.353$$

```
> boxplot(birthweight,ylab="birthweight (g)")
> t.test(x=birthweight, mu=3450, alternative="less")
```

One Sample t-test

```
data: birthweight
t = -2.3529, df = 99, p-value = 0.0103
alternative hypothesis: true mean is less than 3450
95 percent confidence interval:
 -Inf 3420.567
sample estimates:
mean of x
 3350
```

