PSYCH 710

Review of Statistical Inference

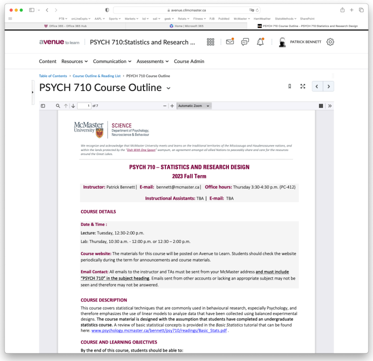
Overview

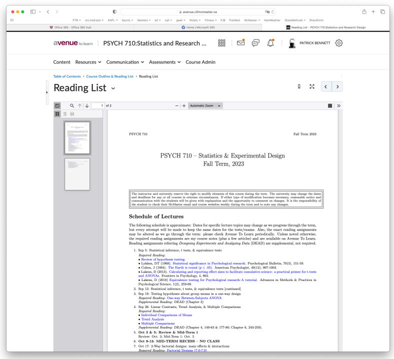
- Course Information
- What is statistics?
- Ways of collecting data
- Modes of statistical analysis
- Sampling distributions & parameter estimation
- z tests & t tests

Course Management

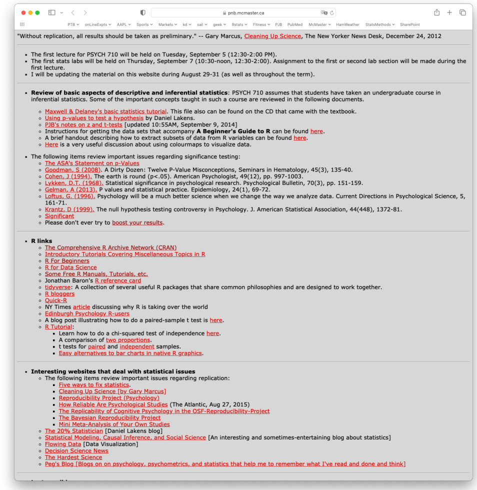
- All lecture materials will be available on Avenue 2 Learn
  - reading assignments
  - lecture slides
  - labs & homework assignments
- Students are expected to install and use R
  - labs & exams will be in Psychology computer cluster (PC-154)
  - you may use your own laptop
- Labs, homework, & exams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

Avenue To Learn

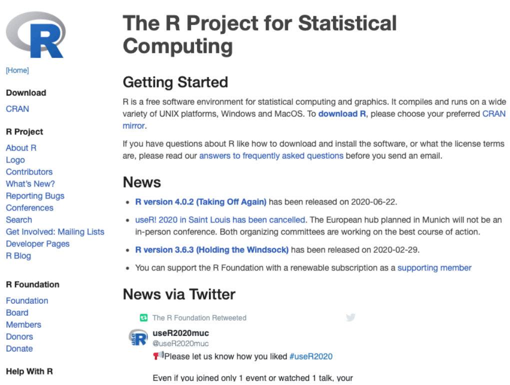




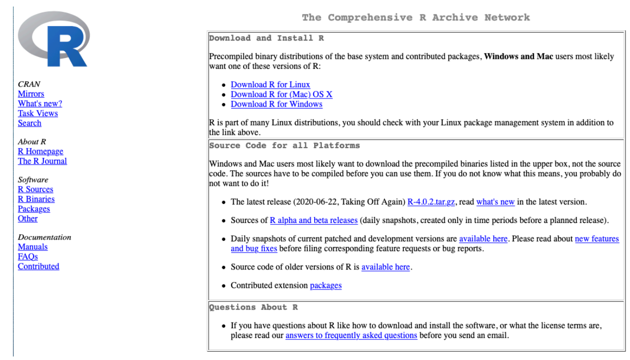
Avenue To Learn



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

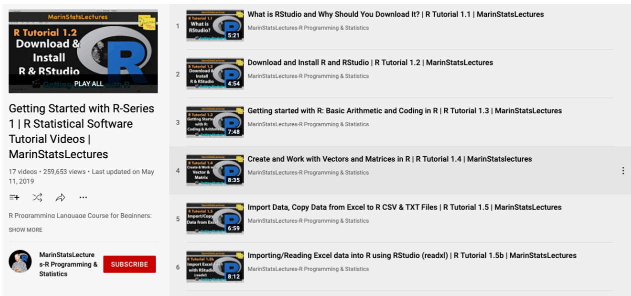


CRAN
https://cran.r-project.org



Video Tutorials on Using R

www.youtube.com/playlist?list=PLqzoL9-eJTNARFXxgwbqGo56NtbJnB37A



Video Tutorials on Using R

www.youtube.com/playlist?list=PLqzoL9-eJTNARFXxgwbqGo56NtbJnB37A

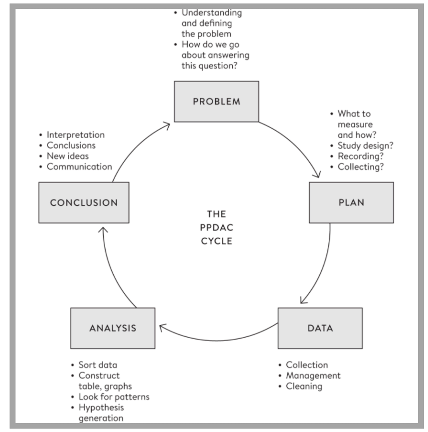
Review of Statistical Inference

What is statistics?

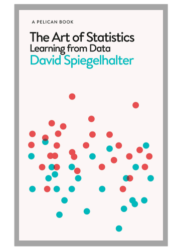
Role of statistics in research

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

Role of statistics in research (PPDAC)



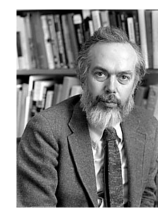

David Spiegelhalter



Role of statistics in research

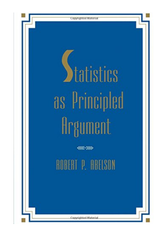
- Statistical methods help us to collect, organize, summarize, analyze, interpret, & present data
- "The role of statistics is not to discover truth. The role of statistics is to resolve disagreements between people." - Milton Friedman

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



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)

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

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

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

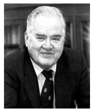
Modes of Statistical Analysis

Descriptive vs. Inferential
Exploratory vs. Confirmatory

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]

Exploratory vs Confirmatory Analyses

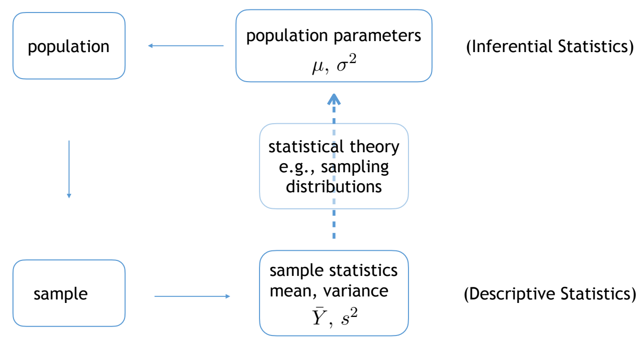


John Tukey

- 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
- <u>Important not to confuse two types of analyses</u>
  - replication crisis in Psychology related to confusion about two types of research

Sampling Distributions & Point Estimators

Descriptive & Inferential Statistics



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?

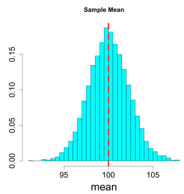
Sample Statistics vs. Population Parameters

- We **<u>can</u>** use sample statistics to estimate population parameters
  - The sample mean, $\bar{Y}$, is an unbiased estimate of population mean, μ
  - The sample <u>variance</u>, s^2, is an unbiased estimate of population <u>variance</u>, 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

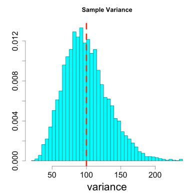
Sampling Distributions of Mean and SD

distributions of statistics
for 5000 samples (n=20)

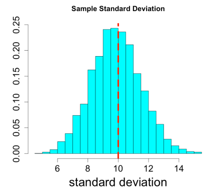
population mean = 100
population variance = 100
population sd = 10



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



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



mean = 9.87
variance = 2.65

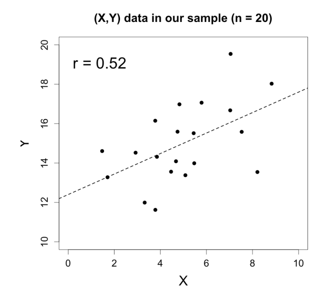
Interval Estimators & Confidence Intervals

Confidence Interval

- 95% Confidence Interval
  - an <u>interval estimate</u> 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 <u>percentile bootstrap</u> method
    - N.B. the method is not as important as understanding the meaning of the $CI_{95\%}$

calculating 95% confidence interval for ρ

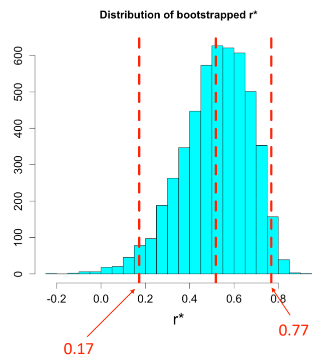
- our sample *r* = 0.52
- what is $CI_{95\%}$ for ρ?
- percentile bootstrap method:
  - uses (X,Y) sample as estimate of (X,Y) population
  - calculate r* on bootstrapped samples:
    - randomly select 20 (X,Y) pairs **<u>from the data</u>**
    - calculate r for each bootstrapped sample (r*)
    - repeat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



distribution of bootstrapped r* values

calculate r* for many <u>bootstrapped samples</u> of data (n=20)

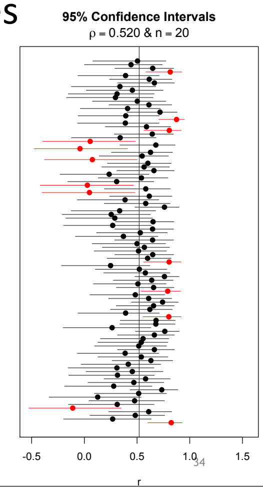
- (X,Y) data is estimate of (X,Y) population
- create bootstrapped sample:
  - randomly select 20 (X,Y) pairs from data
  - calculate r* for each bootstrapped sample (r*)
  - repeat MANY times
  - display histogram of r*
- identify range that contains 95% of r*
  - range is PERCENTILE BOOTSTRAP estimate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



distribution of bootstrapped r* values

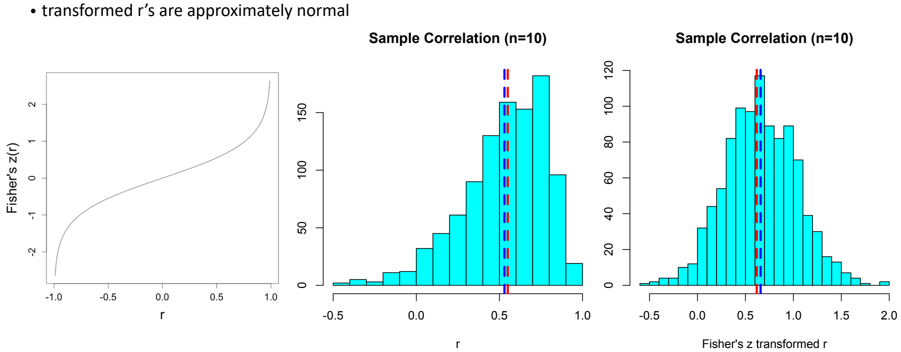
calculate r* for many <u>bootstrapped samples</u> of data (n=20)

- (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



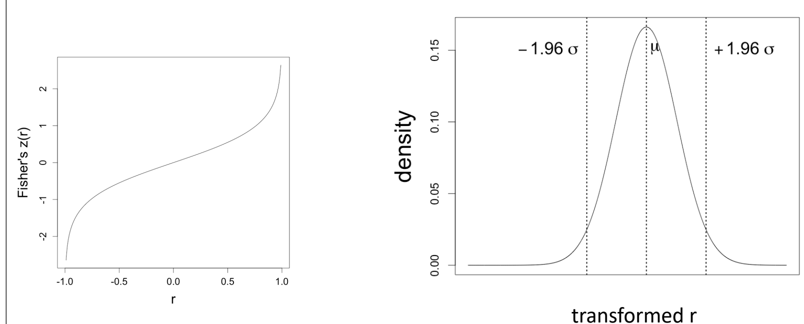
Confidence Interval of r

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



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%



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-951445b9cb2d#

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

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

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

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

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:

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

. . .

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
```


z test of single observation

z test

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

Density Functions & Probability

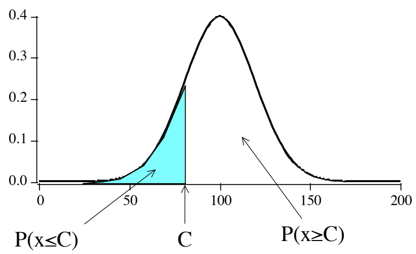


Figure 2: The probability of randomly selecting a value of x that is $\leq C$ – i.e., $P(x \leq C)$ – corresponds to the area under the probability density function that is to the left of C. $P(x \geq C)$ equals the area under the curve that is to the right of C.

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

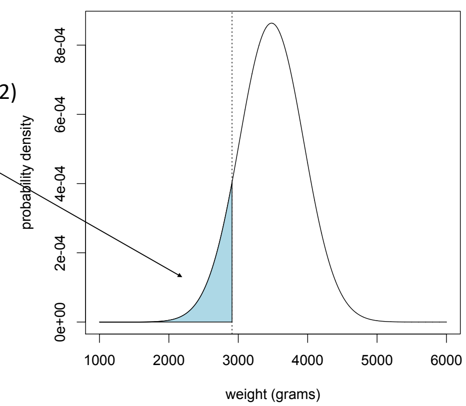
z test

- 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?

z test

p(weight < 2910)
= pnorm(2910, mean=3480, sd=462)
= 0.109

p(z < -1.23)
= pnorm(-1.23,mean=0,sd=1)
= 0.109



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

z test for a group mean

z test for means

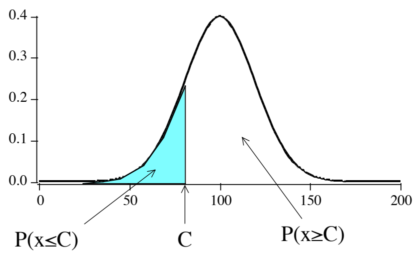
$$z = \frac{(\bar{Y} - \mu_{\bar{Y}})}{\sigma_{\bar{Y}}}$$

- 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?

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

Density Functions & Probability



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 (H0) is:
  - sample is drawn from population with parameters that are identical for Caucasian birth weights
  - distribution of BIRTHWEIGHTS: mean = u = 3480; sd = σ = 462; distribution = NORMAL
  - distribution of MEANS: mean = 3480; **sd = σ/sqrt(n)**; distribution = NORMAL
- when H0 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) = pnorm(-2.81,0,1) = 0.0025
- if sample was drawn from population of Caucasian birth weights, then group mean is unusually low

t test for a group mean

why use t instead of z?

Effect of using <u>estimate</u> of σ

- z is defined with KNOWN population μ and σ
- only source of variation in z is sampling error of mean
- using <u>estimate</u> of σ introduces <u>another</u> 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

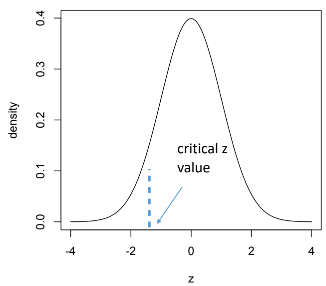
William Sealy Gosset (aka Student)

- 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 **<u>biased</u>** estimate of population standard deviation
  - sample SD <u>underestimates</u> population SD particularly for small sample sizes
- Discovered that **<u>using estimates of σ lead to extreme values of z more frequently than predicted</u>** by statistical theory

Effect of inflating z score

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

- calculating z with <u>estimated</u> σ inflates z scores
- extreme z values **<u>occur more frequently</u>** than expected when H0 is True
- what effect does this have on our evaluation of H0?
  - produces more Type I errors than expected



Critical z = -1.645

p(z ≤ -1.645 | H0) = 0.05

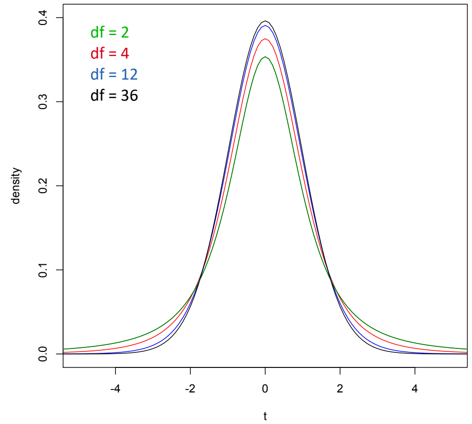
Effect of using estimate of population variation

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
- **<u>t approximately normal for df ≳ 35</u>**



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 <u>not</u> follow z distribution
  - using "z" increases Type I error rate
- However, standardized score <u>DOES</u> follow a *t* distribution
- Therefore, our estimated "z" actually is a t statistic
- so <u>we use critical values of *t*</u>, not *z*, to evaluate null hypothesis

t test of sample mean

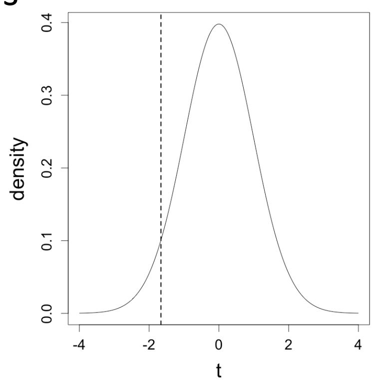
t test for means

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

- 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
  - u = 3480; σ = ???; distribution's shape = ??? [we will assume it is NORMAL]
- is our sample mean of 3350 g unusually low?

t test for means

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



When H0 is true, t will follow t distribution with 100-1=99 degrees of freedom.
This t distribution is <u>very</u> similar to a standard normal discribution.
We expect to get a sample t below the t=-1.66 approximately 5% of the time.

t test for means

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

- population mean is known (3480 g) but variance is <u>unknown</u>
- Sample: mean = 3350 g; standard deviation = 425 g; n = 100
- H0: our sample was drawn from typical population
  - assuming H0 is true, is our sample mean unusually low?
- convert observed mean to a t score:
- compare to critical value of t ($t_{critical}$ = -1.66)
- observed t is more extreme than $t_{critical}$
  - assuming H0 is true, our mean is unusally 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
```


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="birtweight (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
```