T Confidence Intervals

- In the previous, we discussed creating a confidence interval using the CLT
  - They took the form $Est \pm ZQ \times SE_{Est}$
- In this lecture, we discuss some methods for small samples, notably Gosset's t distribution and t confidence intervals
  - They are of the form $Est \pm TQ \times SE_{Est}$
- These are some of the handiest of intervals
- If you want a rule between whether to use a t interval or normal interval, just always use the t interval
- We'll cover the one and two group versions

Gosset's t Distribution

- Invented by William Gosset (under the pseudonym "Student") in 1908
- Has thicker tails than the normal
- Is indexed by a degrees of freedom; gets more like a standard normal as df gets larger
- It assumes that the underlying data are iid Gaussian with the result that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

  follows Gosset's t distribution with $n - 1$ degrees of freedom
- (If we replaced s by σ the statistic would be exactly standard normal)
- Interval is $\bar{X} \pm t_{n-1} S/\sqrt{n}$ where t_{n-1} is the relevant quantile

Code

```
k <- 1000
xvals <- seq(-5, 5, length = k)
myplot <- function(df){
  d <- data.frame(y = c(dnorm(xvals), dt(xvals, df)),
                  x = xvals,
                  dist = factor(rep(c("Normal", "T"), c(k,k))))
  g <- ggplot(d, aes(x = x, y = y))
  g <- g + geom_line(size = 2, aes(colour = dist))
  g
}
manipulate(myplot(mu), mu = slider(1, 20, step = 1))
```

```
pvals <- seq(.5, .99, by = .01)
myplot2 <- function(df){
  d <- data.frame(n= qnorm(pvals),t=qt(pvals, df),
                  p = pvals)
  g <- ggplot(d, aes(x= n, y = t))
  g <- g + geom_abline(size = 2, col = "lightblue")
  g <- g + geom_line(size = 2, col = "black")
  g <- g + geom_vline(xintercept = qnorm(0.975))
  g <- g + geom_hline(yintercept = qt(0.975, df))
  g
}
manipulate(myplot2(df), df = slider(1, 20, step = 1))
```

Note about t-interval

- The t interval technically assumes that the data are iid normal, though it is robust to this assumption
- It works well whenever the distribution of the data is roughly symmetric and mound shaped
- Paired observations are often analyzed using the t interval by taking differences
- For large degrees of freedom, t quantiles become the same as standard normal quantiles; therefore this interval converges to the same interval as the CLT yielded
- For skewed distributions, the spirit of the t interval assumptions are violated
  - Also, for skewed distributions, it doesn't make a lot of sense to center the interval at the mean
  - In this case, consider taking logs or using a different summary like the median
- For highly discrete data, like binary, other intervals are available

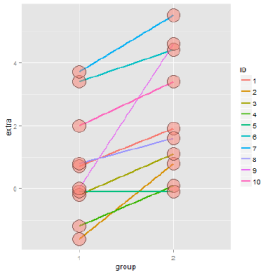
Sleep Data

In R typing `data(sleep)` brings up the sleep data originally analyzed in Gosset's Biometrika paper, which shows the increase in hours for 10 patients on two soporific drugs. R treats the data as two groups rather than paired.

Data

```
data(sleep)
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```



Results

```
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10
```

```
mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n)
t.test(difference)
t.test(g2, g1, paired = TRUE)
t.test(extra ~ I(relevel(group, 2)), paired = TRUE, data = sleep)
```

```
##         [,1] [,2]
## [1,] 0.7001 2.46
## [2,] 0.7001 2.46
## [3,] 0.7001 2.46
## [4,] 0.7001 2.46
```

Independent Group t Confidence Intervals

- Suppose that we want to compare the mean blood pressure between two groups in a randomized trial; those who received the treatment to those who received a placebo
- We cannot use the paired t test because the groups are independent and may have different sample sizes
- We now present methods for comparing independent groups

Confidence Interval

- Therefore a $(1-\alpha)\times 100\%$ confidence interval for $\mu_y - \mu_x$ is

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

- The pooled variance estimator is

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2)$$

- Remember this interval is assuming a constant variance across the two groups
- If there is some doubt, assume a different variance per group, which we will discuss later

Example

Based on Rosner, Fundamentals of Biostatistics

(Really a very good reference book)

- Comparing SBP for 8 oral contraceptive users versus 21 controls
- $\bar{X}_{OC} = 132.86$ mmHg with $s_{OC} = 15.34$ mmHg
- $\bar{X}_{C} = 127.44$ mmHg with $s_{C} = 18.23$ mmHg
- Pooled variance estimate

```
sp <- sqrt((7 * 15.34^2 + 20 * 18.23^2) / (8 + 21 - 2))
132.86 - 127.44 + c(-1, 1) * qt(.975, 27) * sp * (1 / 8 + 1 / 21)^.5
```

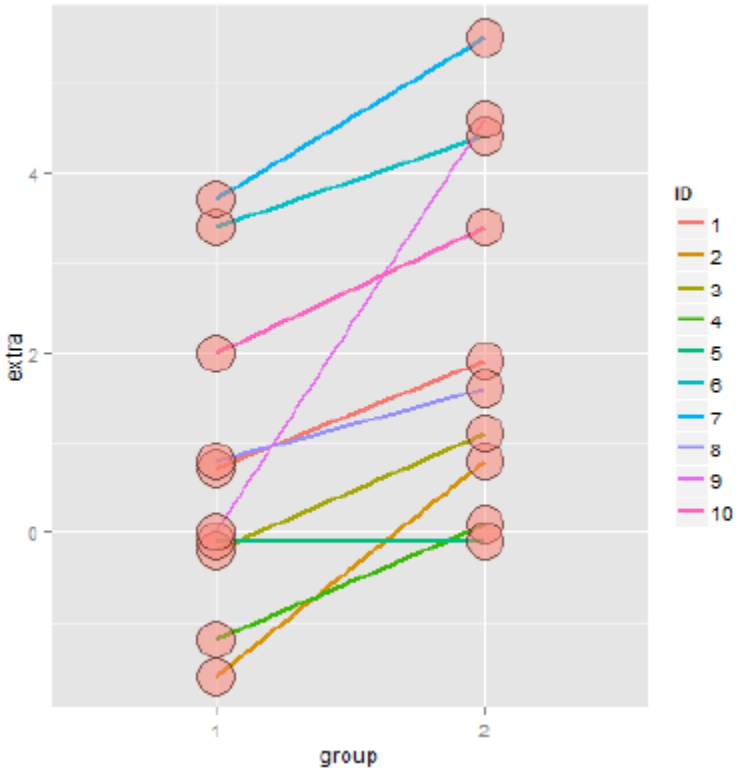
```
## [1] -9.521 20.361
```

Mistakenly Treating the Sleep Data as Grouped

```
n1 <- length(g1); n2 <- length(g2)
sp <- sqrt( ((n1 - 1) * sd(x1)^2 + (n2-1) * sd(x2)^2) / (n1 + n2-2))
md <- mean(g2) - mean(g1)
semd <- sp * sqrt(1 / n1 + 1/n2)
rbind(
md + c(-1, 1) * qt(.975, n1 + n2 - 2) * semd,
t.test(g2, g1, paired = FALSE, var.equal = TRUE)$conf,
t.test(g2, g1, paired = TRUE)$conf
)
```

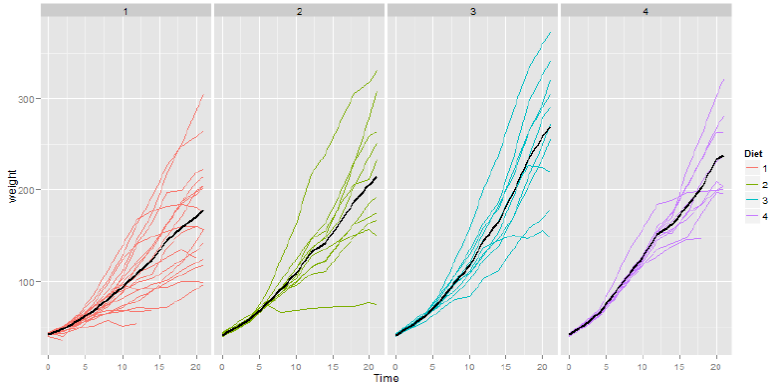
```
##         [,1]  [,2]
## [1,] -0.2039 3.364
## [2,] -0.2039 3.364
## [3,]  0.7001 2.460
```

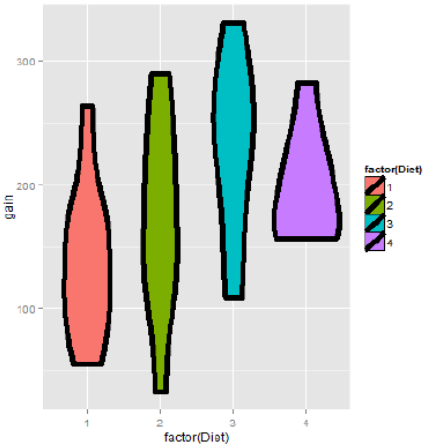
Grouped vs. Independent



ChickWeigh data in R

```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```





A t Interval

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
rbind(
t.test(gain ~ Diet, paired = FALSE, var.equal = TRUE, data = wideCW14)$conf,
t.test(gain ~ Diet, paired = FALSE, var.equal = FALSE, data = wideCW14)$conf
)
```

```
##         [,1]   [,2]
## [1,] -108.1 -14.81
## [2,] -104.7 -18.30
```

Unequal Variances

- Under unequal variances

$$\bar{Y} - \bar{X} \pm t_{df} \times \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^{1/2}$$

  where t_{df} is calculated with degrees of freedom

$$df = \frac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y - 1)}$$

  will be approximately a 95% interval

- This works really well
  - So when in doubt, just assume unequal variances

Example

- Comparing SBP for 8 oral contraceptive users versus 21 controls
- $\bar{X}_{OC} = 132.86$ mmHg with $s_{OC} = 15.34$ mmHg
- $\bar{X}_{C} = 127.44$ mmHg with $s_{C} = 18.23$ mmHg
- $df = 15.04$, $t_{15.04,.975} = 2.13$
- Interval

$$132.86 - 127.44 \pm 2.13\left(\frac{15.34^2}{8} + \frac{18.23^2}{21}\right)^{1/2} = [-8.91, 19.75]$$

- In R, `t.test(..., var.equal = FALSE)`

Comparing other Kinds of Data

- For binomial data, there's lots of ways to compare two groups
  - Relative risk, risk difference, odds ratio.
  - Chi-squared tests, normal approximations, exact tests.
- For count data, there's also Chi-squared tests and exact tests.
- We'll leave the discussions for comparing groups of data for binary and count data until covering glms in the regression class.
- In addition, Mathematical Biostatistics Boot Camp 2 covers many special cases relevant to biostatistics.

Hypothesis Testing

- Hypothesis testing is concerned with making decisions using data
- A null hypothesis is specified that represents the status quo, usually labeled H_0
- The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis

Example

- A respiratory disturbance index of more than 30 events / hour, say, is considered evidence of severe sleep disordered breathing (SDB).
- Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events / hour with a standard deviation of 10 events / hour.
- We might want to test the hypothesis that
  - $H_0 : \mu = 30$
  - $H_a : \mu > 30$
  - where μ is the population mean RDI.
- The alternative hypotheses are typically of the form $<$, $>$ or $\neq$
- Note that there are four possible outcomes of our statistical decision process

TRUTH	DECIDE	RESULT
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

Discussion

- Consider a court of law; the null hypothesis is that the defendant is innocent
- We require a standard on the available evidence to reject the null hypothesis (convict)
- If we set a low standard, then we would increase the percentage of innocent people convicted (type I errors); however we would also increase the percentage of guilty people convicted (correctly rejecting the null)
- If we set a high standard, then we increase the the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors)

Example

- Consider our sleep example again
- A reasonable strategy would reject the null hypothesis if $\bar{X}$ was larger than some constant, say C
- Typically, C is chosen so that the probability of a Type I error, α, is $.05$ (or some other relevant constant)
- α = Type I error rate = Probability of rejecting the null hypothesis when, in fact, the null hypothesis is correct

- Standard error of the mean $10/\sqrt{100} = 1$
- Under H_0 $\bar{X} \sim N(30, 1)$
- We want to chose C so that the $P(\bar{X} > C; H_0)$ is 5%
- The 95th percentile of a normal distribution is 1.645 standard deviations from the mean
- If $C = 30 + 1 \times 1.645 = 31.645$
  - Then the probability that a $N(30, 1)$ is larger than it is 5%
  - So the rule "Reject H_0 when $\bar{X} \geq 31.645$" has the property that the probability of rejection is 5% when H_0 is true (for the μ_0, σ and n given)

Discussion

- In general we don't convert C back to the original scale
- We would just reject because the Z-score; which is how many standard errors the sample mean is above the hypothesized mean

$$\frac{32 - 30}{10/\sqrt{100}} = 2$$

  is greater than 1.645
- Or, whenever $\sqrt{n}(\bar{X} - \mu_0)/s > Z_{1-\alpha}$

General Rules

- The Z test for $H_0 : \mu = \mu_0$ versus
  - $H_1 : \mu < \mu_0$
  - $H_2 : \mu \neq \mu_0$
  - $H_3 : \mu > \mu_0$
- Test statistic $TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Reject the null hypothesis when
  - $TS \leq Z_\alpha = -Z_{1-\alpha}$
  - $|TS| \geq Z_{1-\alpha/2}$
  - $TS \geq Z_{1-\alpha}$

- We have fixed α to be low, so if we reject H_0 (either our model is wrong) or there is a low probability that we have made an error
- We have not fixed the probability of a type II error, β; therefore we tend to say ``Fail to reject H_0" rather than accepting H_0
- Statistical significance is no the same as scientific significance
- The region of TS values for which you reject H_0 is called the rejection region

- The Z test requires the assumptions of the CLT and for n to be large enough for it to apply
- If n is small, then a Gossett's T test is performed exactly in the same way, with the normal quantiles replaced by the appropriate Student's T quantiles and $n - 1$ df
- The probability of rejecting the null hypothesis when it is false is called *power*
- Power is a used a lot to calculate sample sizes for experiments

Example Reconsidered

- Consider our example again. Suppose that $n = 16$ (rather than 100)
- The statistic

$$\frac{\bar{X} - 30}{s/\sqrt{16}}$$

  follows a T distribution with 15 df under H_0
- Under H_0, the probability that it is larger that the 95th percentile of the T distribution is 5%
- The 95th percentile of the T distribution with 15 df is 1.7531 (obtained via `qt(.95, 15)`)
- So that our test statistic is now $\sqrt{16}(32 - 30)/10 = 0.8$
- We now fail to reject.

Two Sided Test

- Suppose that we would reject the null hypothesis if in fact the mean was too large or too small
- That is, we want to test the alternative $H_a : \mu \neq 30$
- We will reject if the test statistic, 0.8, is either too large or too small
- Then we want the probability of rejecting under the null to be 5%, split equally as 2.5% in the upper tail and 2.5% in the lower tail
- Thus we reject if our test statistic is larger than `qt(.975, 15)` or smaller than `qt(.025, 15)`
  - This is the same as saying: reject if the absolute value of our statistic is larger than `qt(0.975, 15)` = 2.1314
  - So we fail to reject the two sided test as well
  - (If you fail to reject the one sided test, you know that you will fail to reject the two sided)

T Test in R

```
library(UsingR); data(father.son)
t.test(father.son$sheight - father.son$fheight)
```

```
>
>   One Sample t-test
>
> data:  father.son$sheight - father.son$fheight
> t = 11.79, df = 1077, p-value < 2.2e-16
> alternative hypothesis: true mean is not equal to 0
> 95 percent confidence interval:
>  0.831 1.163
> sample estimates:
> mean of x
>     0.997
```

Connections with Confidence Intervals

- Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$
- Take the set of all possible values for which you fail to reject H_0, this set is a $(1 - \alpha)100\%$ confidence interval for μ
- The same works in reverse; if a $(1 - \alpha)100\%$ interval contains μ_0, then we *fail to* reject H_0

Two Group Intervals

- First, now you know how to do two group T tests since we already covered indepedent group T intervals
- Rejection rules are the same
- Test $H_0 : \mu_1 = \mu_2$
- Let's just go through an example

Chickweight Data

Recall that we reformatted this data

```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```

Unequal Variance T Test Comparing Diets

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
t.test(gain ~ Diet, paired = FALSE,
       var.equal = TRUE, data = wideCW14)
```

```
>
>   Two Sample t-test
>
>  data:  gain by Diet
>  t = -2.725, df = 23, p-value = 0.01207
>  alternative hypothesis: true difference in means is not equal to 0
>  95 percent confidence interval:
>   -108.15  -14.81
>  sample estimates:
>  mean in group 1 mean in group 4
>            136.2           197.7
```

Exact Binomial Test

- Recall this problem, *Suppose a friend has* 8 *children,* 7 *of which are girls and none are twins*
- Perform the relevant hypothesis test. $H_0 : p = 0.5$ $H_a : p > 0.5$
  - What is the relevant rejection region so that the probability of rejecting is (less than) 5%?

REJECTION REGION	TYPE I ERROR RATE
[0 : 8]	1
[1 : 8]	0.9961
[2 : 8]	0.9648
[3 : 8]	0.8555
[4 : 8]	0.6367
[5 : 8]	0.3633
[6 : 8]	0.1445
[7 : 8]	0.0352
[8 : 8]	0.0039

- It's impossible to get an exact 5% level test for this case due to the discreteness of the binomial.
  - The closest is the rejection region [7 : 8]
  - Any alpha level lower than 0.0039 is not attainable.
- For larger sample sizes, we could do a normal approximation, but you already knew this.
- Two sided test isn't obvious.
  - Given a way to do two sided tests, we could take the set of values of p_0 for which we fail to reject to get an exact binomial confidence interval (called the Clopper/Pearson interval, BTW)
- For these problems, people always create a P-value (next lecture) rather than computing the rejection region.

P-Values

- Most common measure of "statistical significance"
- Their ubiquity, along with concern over their interpretation and use makes them controversial among statisticians
  - http://warnercnr.colostate.edu/~anderson/thompson1.html
  - Also see *Statistical Evidence: A Likelihood Paradigm* by Richard Royall
  - *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* by Steve Goodman
  - The hilariously titled: *The Earth is Round (p < .05)* by Cohen.
- Some positive comments
  - simply statistics
  - normal deviate
  - Error statistics

What is a P-value?

Idea: Suppose nothing is going on - how unusual is it to see the estimate we got?

Approach:

1. Define the hypothetical distribution of a data summary (statistic) when "nothing is going on" (*null hypothesis*)
2. Calculate the summary/statistic with the data we have (*test statistic*)
3. Compare what we calculated to our hypothetical distribution and see if the value is "extreme" (*p-value*)

- The P-value is the probability under the null hypothesis of obtaining evidence as extreme or more extreme than would be observed by chance alone
- If the P-value is small, then either H_0 is true and we have observed a rare event or H_0 is false
- In our example the T statistic was 0.8.
  - What's the probability of getting a T statistic as large as 0.8?

```
pt(0.8, 15, lower.tail = FALSE)
```

```
## [1] 0.2181
```

- Therefore, the probability of seeing evidence as extreme or more extreme than that actually obtained under H_0 is 0.2181

Attained Significance Level

- Our test statistic was 2 for $H_0 : \mu_0 = 30$ versus $H_a : \mu > 30$.
- Notice that we rejected the one sided test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$, how about 0.001?
- The smallest value for alpha that you still reject the null hypothesis is called the *attained significance level*
- This is equivalent, but philosophically a little different from, the *P-value*

- By reporting a P-value the reader can perform the hypothesis test at whatever α level he or she choses
- If the P-value is less than α you reject the null hypothesis
- For two sided hypothesis test, double the smaller of the two one sided hypothesis test Pvalues

Revisiting an Earlier Example

- Suppose a friend has 8 children, 7 of which are girls and none are twins
- If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

```
choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
```

```
## [1] 0.03516
```

```
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
```

```
## [1] 0.03516
```

Poisson Example

- Suppose that a hospital has an infection rate of 10 infections per 100 person/days at risk (rate of 0.1) during the last monitoring period.
- Assume that an infection rate of 0.05 is an important benchmark.
- Given the model, could the observed rate being larger than 0.05 be attributed to chance?
- Under $H_0 : \lambda = 0.05$ so that $\lambda_0 100 = 5$
- Consider $H_a : \lambda > 0.05$.

```
ppois(9, 5, lower.tail = FALSE)
```

```
## [1] 0.03183
```

Power

- Power is the probability of rejecting the null hypothesis when it is false
- Ergo, power (as its name would suggest) is a good thing; you want more power
- A type II error (a bad thing, as its name would suggest) is failing to reject the null hypothesis when it's false; the probability of a type II error is usually called β
- Note Power $= 1 - \beta$

- Consider our previous example involving RDI
- $H_0 : \mu = 30$ versus $H_a : \mu > 30$
- Then power is

$$P\left(\frac{\bar{X} - 30}{s/\sqrt{n}} > t_{1-\alpha,n-1} \; ; \; \mu = \mu_a\right)$$

- Note that this is a function that depends on the specific value of μ_a!
- Notice as μ_a approaches 30 the power approaches α

Calculating Power for Gaussian Data

- We reject if $\frac{\bar{X}-30}{\sigma/\sqrt{n}} > z_{1-\alpha}$
  - Equivalently if $\bar{X} > 30 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
- Under $H_0 : \bar{X} \sim N(\mu_0, \sigma^2/n)$
- Under $H_a : \bar{X} \sim N(\mu_a, \sigma^2/n)$
- So we want

```
alpha = 0.05
z = qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALSE)
```

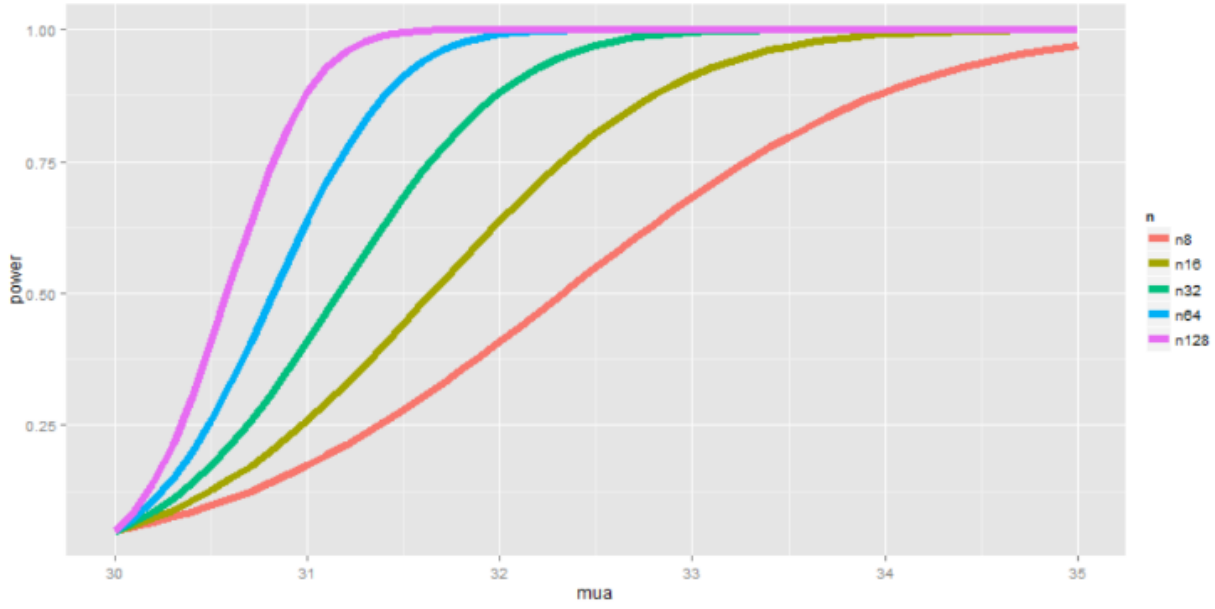
- $\mu_a = 32, \mu_0 = 30, n = 16, \sigma = 4$

```
mu0 = 30
mua = 32
sigma = 4
n = 16
z = qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mu0, sd = sigma/sqrt(n), lower.tail = FALSE)
```

```
## [1] 0.05
```

```
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALSE)
```

```
## [1] 0.6388
```



Graphical Depiction of Power

```
library(manipulate)
mu0 = 30
myplot <- function(sigma, mua, n, alpha) {
    g = ggplot(data.frame(mu = c(27, 36)), aes(x = mu))
    g = g + stat_function(fun = dnorm, geom = "line", args = list(mean = mu0,
        sd = sigma/sqrt(n)), size = 2, col = "red")
    g = g + stat_function(fun = dnorm, geom = "line", args = list(mean = mua,
        sd = sigma/sqrt(n)), size = 2, col = "blue")
    xitc = mu0 + qnorm(1 - alpha) * sigma/sqrt(n)
    g = g + geom_vline(xintercept = xitc, size = 3)
    g
}
manipulate(myplot(sigma, mua, n, alpha), sigma = slider(1, 10, step = 1, initial = 4),
    mua = slider(30, 35, step = 1, initial = 32), n = slider(1, 50, step = 1,
        initial = 16), alpha = slider(0.01, 0.1, step = 0.01, initial = 0.05))
```

Question

- When testing $H_a : \mu > \mu_0$, notice if power is $1 - \beta$, then

$$1 - \beta = P\left(\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \mu = \mu_a\right)$$

- where $\bar{X} \sim N(\mu_a, \sigma^2/n)$
- Unknowns: μ_a, σ, n, β
- Knowns: μ_0, α
- Specify any 3 of the unknowns and you can solve for the remainder

Notes

- The calculation for $H_a : \mu < \mu_0$ is similar
- For $H_a : \mu \neq \mu_0$ calculate the one sided power using $\alpha/2$ (this is only approximately right, it excludes the probability of getting a large TS in the opposite direction of the truth)
- Power goes up as α gets larger
- Power of a one sided test is greater than the power of the associated two sided test
- Power goes up as μ_1 gets further away from μ_0
- Power goes up as n goes up
- Power doesn't need μ_a, σ and n, instead only $\frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma}$
  - The quantity $\frac{\mu_a - \mu_0}{\sigma}$ is called the effect size, the difference in the means in standard deviation units.
  - Being unit free, it has some hope of interpretability across settings

T-test Power

- Consider calculating power for a Gossett's T test for our example
- The power is

$$P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha,n-1} \; ; \; \mu = \mu_a\right)$$

- Calcuting this requires the non-central t distribution.
- `power.t.test` does this very well
  - Omit one of the arguments and it solves for it

Example

```
power.t.test(n = 16, delta = 2/4, sd = 1, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(n = 16, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(n = 16, delta = 100, sd = 200, type = "one.sample", alt = "one.sided")$power
```

```
## [1] 0.604
```

```
power.t.test(power = 0.8, delta = 2/4, sd = 1, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

```
power.t.test(power = 0.8, delta = 2, sd = 4, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

```
power.t.test(power = 0.8, delta = 100, sd = 200, type = "one.sample", alt = "one.sided")$n
```

```
## [1] 26.14
```

Multiple Testing

- Hypothesis testing/significance analysis is commonly overused
- Correcting for multiple testing avoids false positives or discoveries
- Two key components
  - Error measure
  - Correction

Three Eras of Statistics

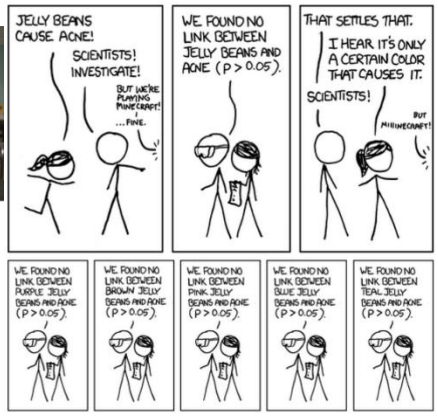
The age of Quetelet and his successors, in which huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?

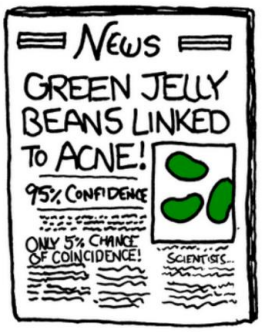
The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who **developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment**. The questions dealt with still tended to be simple Is treatment A better than treatment B?

The era of scientific mass production, in which new technologies typified by the microarray allow a single team of scientists to produce data sets of a size Quetelet would envy. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together; not at all what the classical masters had in mind. Which variables matter among the thousands measured? How do you relate unrelated information?

http://www-stat.stanford.edu/~ckirby/brad/papers/2010LSIexcerpt.pdf

Why correct for multiple tests?





Types of Errors

Suppose you are testing a hypothesis that a parameter β equals zero versus the alternative that it does not equal zero. These are the possible outcomes.

	$\beta = 0$	$\beta \neq 0$	HYPOTHESES
Claim $\beta = 0$	U	T	$m - R$
Claim $\beta \neq 0$	V	S	R
Claims	m_0	$m - m_0$	m

Type I error or false positive (V) Say that the parameter does not equal zero when it does

Type II error or false negative (T) Say that the parameter equals zero when it doesn't

Error Rates

False positive rate - The rate at which false results ($\beta = 0$) are called significant: $E\left[\frac{V}{m_0}\right]$ *

Family wise error rate (FWER) - The probability of at least one false positive $\Pr(V \geq 1)$

False discovery rate (FDR) - The rate at which claims of significance are false $E\left[\frac{V}{R}\right]$

- The false positive rate is closely related to the type I error rate http://en.wikipedia.org/wiki/False_positive_rate

Controlling the False Positive Rate

If P-values are correctly calculated calling all $P < \alpha$ significant will control the false positive rate at level α on average.

Problem: Suppose that you perform 10,000 tests and $\beta = 0$ for all of them.

Suppose that you call all $P < 0.05$ significant.

The expected number of false positives is: $10,000 \times 0.05 = 500$ false positives.

How do we avoid so many false positives?

Controlling Family-wise Error Rate (FWER)

The Bonferroni correction is the oldest multiple testing correction.

Basic idea:

- Suppose you do m tests
- You want to control FWER at level α so $Pr(V \geq 1) < \alpha$
- Calculate P-values normally
- Set $\alpha_{fwer} = \alpha/m$
- Call all P-values less than α_{fwer} significant

Pros: Easy to calculate, conservative **Cons**: May be very conservative

Controlling False Discovery Rate (FDR)

This is the most popular correction when performing *lots* of tests say in genomics, imaging, astronomy, or other signal-processing disciplines.

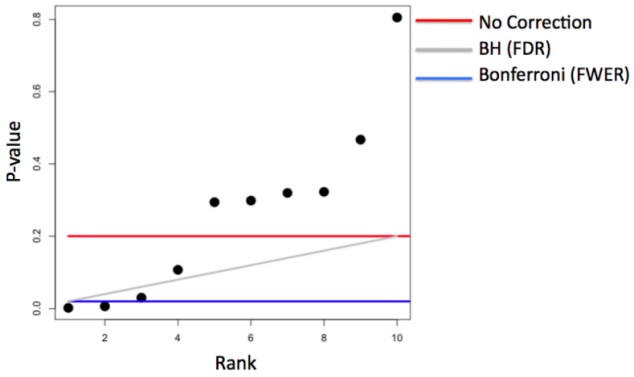
Basic idea:

- Suppose you do m tests
- You want to control FDR at level α so $E\left[\frac{V}{R}\right]$
- Calculate P-values normally
- Order the P-values from smallest to largest $P_{(1)}, \ldots, P_{(m)}$
- Call any $P_{(i)} \leq \alpha \times \frac{i}{m}$ significant

Pros: Still pretty easy to calculate, less conservative (maybe much less)

Cons: Allows for more false positives, may behave strangely under dependence

Example with 10 P-values



Controlling all error rates at $\alpha = 0.20$

Adjusted P-values

- One approach is to adjust the threshold α
- A different approach is to calculate "adjusted p-values"
- They *are not p-values* anymore
- But they can be used directly without adjusting α

Example:

- Suppose P-values are $P_1, \ldots, P_m$
- You could adjust them by taking $P_i^{fwer} = \max m \times P_i, 1$ for each P-value.
- Then if you call all $P_i^{fwer} < \alpha$ significant you will control the FWER.

Case Study 1: No True Positives

```
set.seed(1010093)
pValues <- rep(NA, 1000)
for (i in 1:1000) {
    y <- rnorm(20)
    x <- rnorm(20)
    pValues[i] <- summary(lm(y ~ x))$coeff[2, 4]
}

# Controls false positive rate
sum(pValues < 0.05)
```

```
## [1] 51
```

```
# Controls FWER
sum(p.adjust(pValues, method = "bonferroni") < 0.05)
```

```
## [1] 0
```

```
# Controls FDR
sum(p.adjust(pValues, method = "BH") < 0.05)
```

```
## [1] 0
```

Case Study 2: 50% True Positives

```
set.seed(1010093)
pValues <- rep(NA, 1000)
for (i in 1:1000) {
    x <- rnorm(20)
    # First 500 beta=0, last 500 beta=2
    if (i <= 500) {
        y <- rnorm(20)
    } else {
        y <- rnorm(20, mean = 2 * x)
    }
    pValues[i] <- summary(lm(y ~ x))$coeff[2, 4]
}
trueStatus <- rep(c("zero", "not zero"), each = 500)
table(pValues < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  476
##   TRUE       500   24
```

```
# Controls FWER
table(p.adjust(pValues, method = "bonferroni") < 0.05, trueStatus)
```

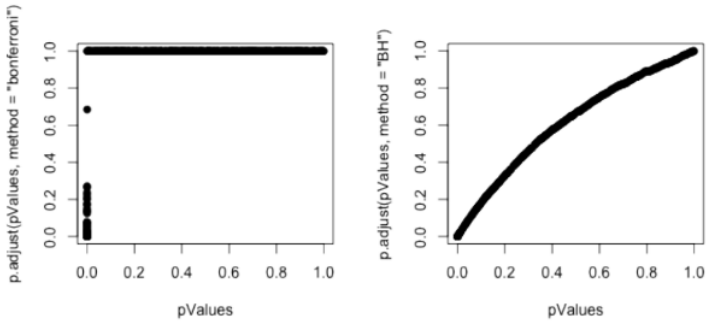
```
##        trueStatus
##         not zero zero
##   FALSE       23  500
##   TRUE       477    0
```

```
# Controls FDR
table(p.adjust(pValues, method = "BH") < 0.05, trueStatus)
```

```
##        trueStatus
##         not zero zero
##   FALSE        0  487
##   TRUE       500   13
```

P-values Versus Adjusted Values

```
par(mfrow = c(1, 2))
plot(pValues, p.adjust(pValues, method = "bonferroni"), pch = 19)
plot(pValues, p.adjust(pValues, method = "BH"), pch = 19)
```



Notes and Resources

Notes:

- Multiple testing is an entire subfield
- A basic Bonferroni/BH correction is usually enough
- If there is strong dependence between tests there may be problems
  - Consider method="BY"

Further resources:

- Multiple testing procedures with applications to genomics
- Statistical significance for genome-wide studies
- Introduction to multiple testing

Resampled Interface
The Jackknife

- The jackknife is a tool for estimating standard errors and the bias of estimators
- As its name suggests, the jackknife is a small, handy tool; in contrast to the bootstrap, which is then the moral equivalent of a giant workshop full of tools
- Both the jackknife and the bootstrap involve *resampling* data; that is, repeatedly creating new data sets from the original data
- The jackknife deletes each observation and calculates an estimate based on the remaining $n-1$ of them
- It uses this collection of estimates to do things like estimate the bias and the standard error
- Note that estimating the bias and having a standard error are not needed for things like sample means, which we know are unbiased estimates of population means and what their standard errors are
- We'll consider the jackknife for univariate data
- Let $X_1, \ldots, X_n$ be a collection of data used to estimate a parameter θ
- Let $\hat{\theta}$ be the estimate based on the full data set
- Let $\hat{\theta}_i$ be the estimate of θ obtained by *deleting observation* i
- Let $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i$
- Then, the jackknife estimate of the bias is

$$(n-1)\left(\bar{\theta} - \hat{\theta}\right)$$

(how far the average delete-one estimate is from the actual estimate)

- The jackknife estimate of the standard error is

$$\left[\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2\right]^{1/2}$$

(the deviance of the delete-one estimates from the average delete-one estimate)

Example

We want to estimate the bias and standard error of the median

```
library(UsingR)
data(father.son)
x <- father.son$sheight
n <- length(x)
theta <- median(x)
jk <- sapply(1:n, function(i) median(x[-i]))
thetaBar <- mean(jk)
biasEst <- (n - 1) * (thetaBar - theta)
seEst <- sqrt((n - 1) * mean((jk - thetaBar)^2))
```

```
c(biasEst, seEst)
```

```
## [1] 0.0000 0.1014
```

```
library(bootstrap)
temp <- jackknife(x, median)
c(temp$jack.bias, temp$jack.se)
```

```
## [1] 0.0000 0.1014
```

- Both methods (of course) yield an estimated bias of 0 and a se of 0.1014
- Odd little fact: the jackknife estimate of the bias for the median is always 0 when the number of observations is even
- It has been shown that the jackknife is a linear approximation to the bootstrap
- Generally do not use the jackknife for sample quantiles like the median; as it has been shown to have some poor properties

Pseudo Observations

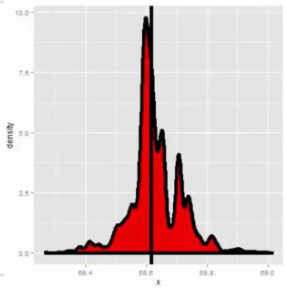
- Another interesting way to think about the jackknife uses pseudo observations
- Let

$$\text{Pseudo Obs} = n\hat{\theta} - (n-1)\hat{\theta}_i$$

- Think of these as ``whatever observation i contributes to the estimate of θ"
- Note when $\hat{\theta}$ is the sample mean, the pseudo observations are the data themselves
- Then the sample standard error of these observations is the previous jackknife estimated standard error.
- The mean of these observations is a bias-corrected estimate of θ

Bootstrapping Example

```
library(UsingR)
data(father.son)
x <- father.son$sheight
n <- length(x)
B <- 10000
resamples <- matrix(sample(x, n * B, replace = TRUE), B, n)
resampledMedians <- apply(resamples, 1, median)
```



The general procedure follows by first simulating complete data sets from the observed data with replacement

- This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution

Calculate the statistic for each simulated data set

Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error

Nonparametric Bootstrap Alg. Ex.

Bootstrap procedure for calculating confidence interval for the median from a data set of n observations

i. Sample n observations **with replacement** from the observed data resulting in one simulated complete data set

ii. Take the median of the simulated data set

iii. Repeat these two steps B times, resulting in B simulated medians

iv. These medians are approximately drawn from the sampling distribution of the median of n observations; therefore we can

iv. These medians are approximately drawn from the sampling distribution of the median of n observations; therefore we can

- Draw a histogram of them
- Calculate their standard deviation to estimate the standard error of the median
- Take the 2.5^{th} and 97.5^{th} percentiles as a confidence interval for the median

Example Code

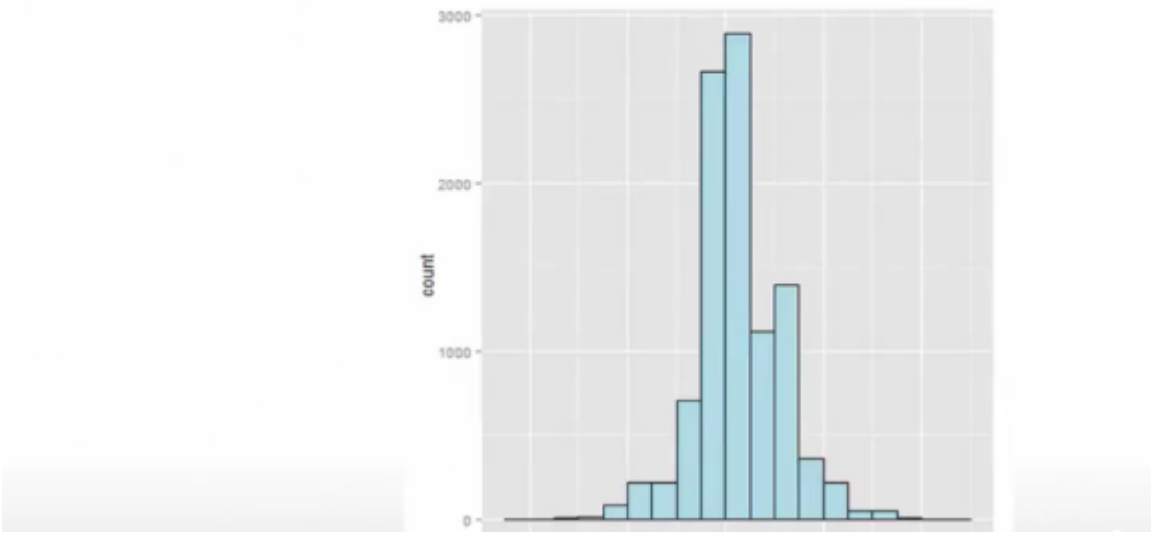
```r
B <- 10000
resamples <- matrix(sample(x, n * B, replace = TRUE), B, n)
medians <- apply(resamples, 1, median)
sd(medians)
```

```
## [1] 0.08473
```

```r
quantile(medians, c(0.025, 0.975))
```

```
##  2.5% 97.5%
## 68.43 68.82
```

```r
g = ggplot(data.frame(medians = medians), aes(x = medians))
g = g + geom_histogram(color = "black", fill = "lightblue", binwidth = 0.05)
g
```

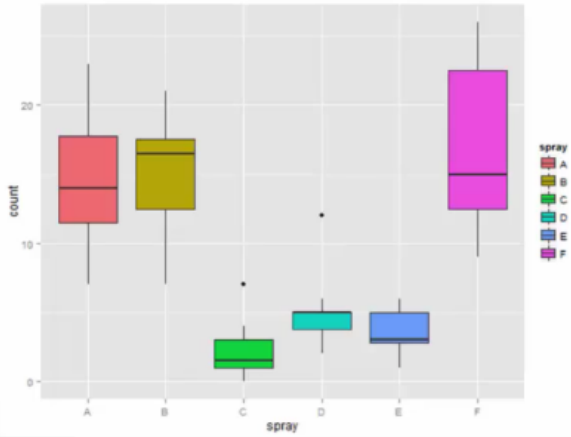


Notes on Bootstrap

- The bootstrap is non-parametric
- Better percentile bootstrap confidence intervals correct for bias
- There are lots of variations on bootstrap procedures; the book "An Introduction to the Bootstrap"" by Efron and Tibshirani is a great place to start for both bootstrap and jackknife information

Group Comparisons

- Consider comparing two independent groups.
- Example, comparing sprays B and C



Permutation Tests

- Consider the null hypothesis that the distribution of the observations from each group is the same
- Then, the group labels are irrelevant
- Consider a data frome with count and spray
- Permute the spray (group) labels
- Recalculate the statistic
  - Mean difference in counts
  - Geometric means
  - T statistic
- Calculate the percentage of simulations where the simulated statistic was more extreme (toward the alternative) than the observed

Variations on Permutation Testing

DATA TYPE	STATISTIC	TEST NAME
Ranks	rank sum	rank sum test
Binary	hypergeometric prob	Fisher's exact test
Raw data		ordinary permutation test

- Also, so-called *randomization tests* are exactly permutation tests, with a different motivation.
- For matched data, one can randomize the signs
  - For ranks, this results in the signed rank test
- Permutation strategies work for regression as well
  - Permuting a regressor of interest
- Permutation tests work very well in multivariate settings

Permutation Test B vs. C

```
subdata <- InsectSprays[InsectSprays$spray %in% c("B", "C"),]
y <- subdata$count
group <- as.character(subdata$spray)
testStat <- function(w, g) mean(w[g == "B"]) - mean(w[g == "C"])
observedStat <- testStat(y, group)
permutations <- sapply(1 : 10000, function(i) testStat(y, sample(group)))
observedStat
```

```
## [1] 13.25
```

```
mean(permutations > observedStat)
```

```
## [1] 0
```