Confidence Intervals

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

February 12, 2014

What is this Presentation?

- Terminology review
- The Idea of a CI
- Proportions
- Means
- Etc

What do you really need to learn?

- The big idea: we make estimates, try to summarize our uncertainty about them.
- The Conf Interval idea presumes we can
  - imagine a sampling distribution
  - find a way, using only one sample, get estimate of how uncertain we are
- This can be tricky in some cases, but we try to understand the important cases clearly (and hope we can read a manual when we come to unusual ones)

Recall Terminology:

Parameter: θ is a "parameter", a "true value" that governs a "data generating process." It is the characteristic of the thing from which we draw observations, which in statistics is often called "the population". Because that is confusing/value laden, I avoid "population" terminology.

Parameter Estimate: $\hat{\theta}$ is a number that gets calculated from sample data. Hopefully, it is

- consistent (reminder from last lecture).

Sampling Distribution: the assumed probability model for $\hat{\theta}$. If a particular theory about θ is correct, what would be the PDF of $\hat{\theta}$?

A Sampling Distributions is characterized by an Expected Value and Variance (as are all random variables).

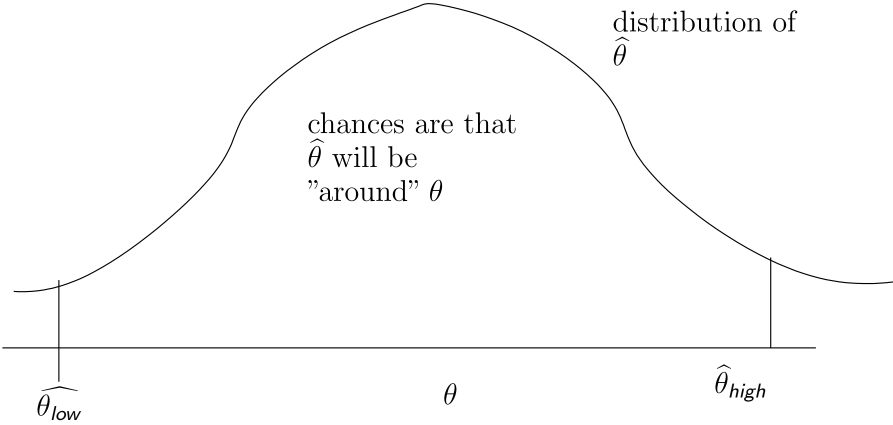
Standard Error: From one sample, estimate the standard deviation of $\hat{\theta}$ (How much $\hat{\theta}$ would vary if we collected a lot of estimates). Recall the silly notation, $\sqrt{\widehat{Var(\hat{\theta})}}$, The estimate of the uncertainty of an estimate.

Today's Focus: Confidence Interval

- General idea: We know that estimates from samples are not exactly equal to the "true" parameters we want to estimate
- Ever watch CNN report that "41% of Americans favor XYZ, plus-or-minus 3%"

Sampling Dist.

Suppose you know that the Sampling Dist is like so:



This was selected from the elaborate collection of ugly distributions, a freely available library that I can share to you any time you like :).

Outline

Define Confidence Interval

- $\hat{\theta}$ is a estimate from a sample, a value that would fluctuate from sample-to-sample
- **Confidence Interval**: From one estimate $\hat{\theta}$, construct a range $[\hat{\theta}_{low}, \hat{\theta}_{high}]$ that we think is likely to contain the truth.
- We decide "how likely" it must be that the truth is in there, then we construct the CI. Common to use 95%.
- A 95% Confidence Interval would have 2 meanings
  - 1. Repeated Sampling: 95% of sample estimates would fall into $[\hat{\theta}_{low}, \ \hat{\theta}_{high}]$
  - 2. Degree of Belief: The probability is 0.95 that θ is in $[\hat{\theta}_{low}, \ \hat{\theta}_{high}]$

CI: The First Interpretation: Repeated Sampling

- If you knew the sampling distribution, you could get a math genius to figure out the range.

$$Prob(\widehat{\theta_{low}} < \hat{\theta} < \widehat{\theta_{high}}) \tag{1}$$

  This pre-supposes you know the "true θ" and the PDF of $\hat{\theta}$. (And that you know a math genius.)
- One custom is to pick the low and high edges so that

$$Prob(\widehat{\theta_{low}} < \hat{\theta} < \widehat{\theta_{high}}) = 0.95 \tag{2}$$

  If we repeated this experiment over and over, then the probability that the estimate will be between $\widehat{\theta_{low}}$ and $\widehat{\theta_{high}}$ is 0.95.
- Repeat: There is a 95% chance that a random sample estimate will lie between the two edges.
- The "p-value" in statistics is the part that is outside of that range. Here, $p = 0.05$.
- "p-value" sometimes referred to as α, or *alpha level*.

CI: Second Interpretation: The Degree of Belief

- This is a stronger statement, one I resisted for many years:

Theorem

Construct a CI $[\widehat{\theta_{low}}, \widehat{\theta_{high}}]$ *from one sample. The probability that the true value of* θ *is in that interval is 0.95.*

Work through Verzani's argument

Claim: Given $\hat{\theta}$, there is a 0.95 probability (a 95% chance) that the "true value of θ" is between $\hat{\theta}_{low}$ and $\hat{\theta}_{high}$.

- Think of the low and high edges as plus or minus the true θ:

$$Prob(\theta - \textit{something on the left} < \widehat{\theta} < \theta + \textit{something on the right}) = 0.95 \quad (3)$$

If the Sampling Distribution is Symmetric

- If the sampling distribution is symmetric, we subtract and add the same "something" on either side.

$$Prob(\theta - something < \widehat{\theta} < \theta + something) = 0.95$$

- Subtract θ from each term

$$Prob(-something < \widehat{\theta} - \theta < something) = 0.95$$

- Subtract $\widehat{\theta}$ from each term

$$Prob(-\hat{\theta} - something < -\theta < -\widehat{\theta} + something) = 0.95$$

- Multiply through by -1 and you get ....

The Big Conclusion:

- A Confidence Interval is

$$Prob(\hat{\theta} - something < \theta < \widehat{\theta} + something) = 0.95 \tag{4}$$

- We believe "with 95% confidence" that the true value will lie between two outside edges,

$$[\hat{\theta} - something, \, \hat{\theta} + something]$$

.

- The *something* is the "margin of error"

Outline

The Challenge: Find Way To Calculate CIs

- A CI requires us to know the sampling distribution of $\hat{\theta}$, and then we:
  - "grab" the middle 95%
- Not all CIs are symmetric, but the easiest ones to visualize are symmetric (estimated means, slope coefficients)
- Symmetric CI: $[\hat{\theta}_{low}, \hat{\theta}_{high}] = [\hat{\theta} - something, \hat{\theta} + something]$
- If sampling distribution of $\hat{\theta}$ not symmetric, problem is harder. Will need a formula like

$$[\hat{\theta} - something\ left, \hat{\theta} + something\ right]$$

Every Estimator has its own CI formula

- The challenge of the CI is that there is no universal formula
- For some estimates, we have "known solutions".
- R has a function confint() for some estimators
- Some estimators have no agreed-upon CI.

Many Symmetric CIs have a simple/similar formula

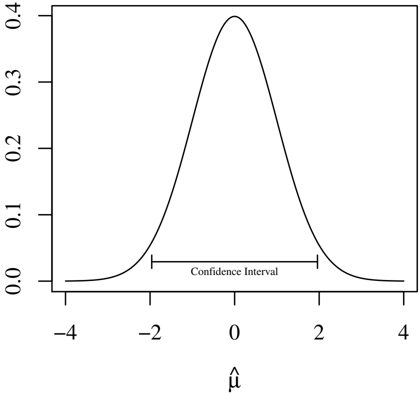
- Put the estimate $\hat{\theta}$ in the center
- Calculate *something* to add and subtract. Generally, it depends on
    1. Standard error of the estimate
    2. Sample size

Outline

If We Knew the Sampling Distribution, life would be easy

- Suppose $\hat{\mu}$ has a sampling distribution that is Normal with variance 1, i.e., $N(\mu, 1)$.
- An observation $\hat{\mu}$ is an unbiased estimator of μ.
- Since $\sigma^2 = 1$, our knowledge of the Normal tells us that μ is very likely in this region

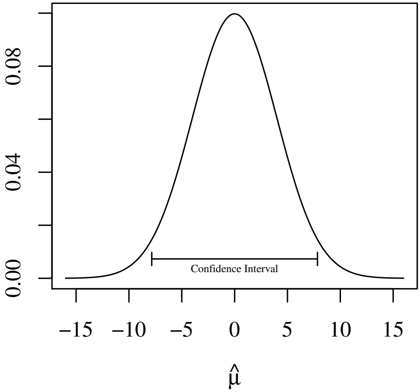
$$Prob(\mu \in [\hat{\mu} - 1.96, \hat{\mu} + 1.96]) = 0.95$$



Suppose σ were 4

- Suppose $\hat{\mu}$is Normal, but with standard deviation $sd(\hat{\mu}) = \sigma = 4$. Then $\hat{\mu} \sim N(0, 4^2)$.
- The 0.95 CI is

$$[\hat{\mu} - 1.96 \cdot 4), \hat{\mu} + 1.96 \cdot 4]$$



How do we know 1.96 is the magic number?

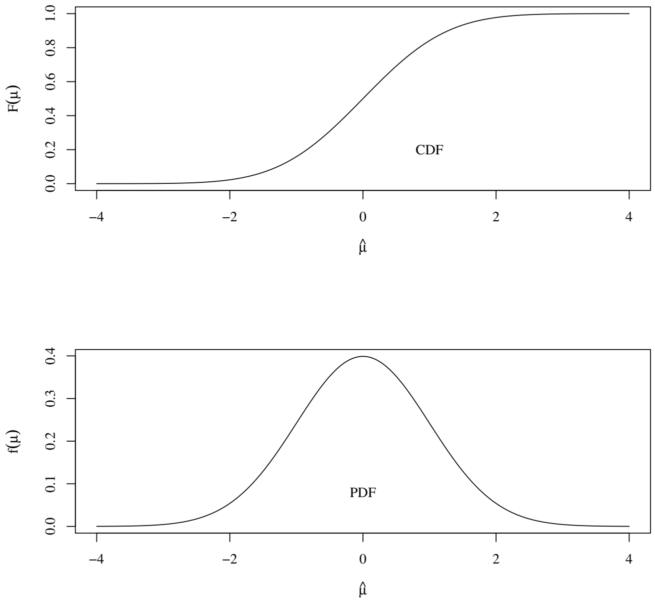
Correct Answer We stipulated that the sampling distribution was Normal. The probability of an outcome below -1.96 is 0.025 and the chance of an outcome greater than 1.96 is 0.025.

Another Correct Answer In the old days, we'd look it up in a stats book that has the table of Normal Probabilities.

Another Correct Answer Today, we ask R, using the qnorm function:

```
> qnorm(0.025, m = 0, sd = 1)
[1] -1.959964
```

The value $-1.959964 \approx -1.96$ is greater than 0.025 of the possible outcomes.


1.0
0.8
0.6
0.4
0.2
0.0
F(μ)
CDF
−4
−2
0
2
4
μ̂
0.4
0.3
0.2
0.1
0.0
f(μ)
PDF
−4
−2
0
2
4
μ̂

Some Example Values

- Some easy to remember values from the Standard Normal are

Examples:	**> qnorm(0.5)** **[1] 0**	**> qnorm(0.05)** **[1] -1.6448**

Some values from the CDF:

$F(-\infty) = 0$	$F(-1.96) = 0.025$	$F(-1.65) = 0.05$	$F(0) = 0.5$
	$F(1.65) = 0.95$	$F(1.96) = 0.975$	$F(\infty) = 1$

- Conclusion: The $\alpha = 0.05$ confidence interval for a estimator that is $N(\mu, 1)$ is

$$(\hat{\mu} - 1.96, \hat{\mu} + 1.96) \tag{5}$$

The Sampling Distribution of the Mean/Std.Err.(Mean)

- Previous supposed I knew σ, the "true" standard deviation of $\hat{\mu}$.
- Now I make the problem more challenging, forcing myself to estimate the mean, and standard error of the mean.
- In the end, we NEVER create a sampling distribution for the mean by itself.
- We DO estimate the sampling distribution of the ratio of the "estimation mean" $(\hat{\mu} - \mu)$ to its standard error.
- Intuition: The CI will be symmetric, $\hat{\mu} \pm$ *something*, using the sampling distribution

Sample Mean

- Collect some observations, $x_1, x_2, x_3, \ldots, x_N$
- The sample mean (call it $\bar{x}$ or $\hat{\mu}$) is an estimate of the "expected value",

$$\textit{sample mean of } x : \bar{x} = \hat{\mu} = \frac{1}{N}\sum x_i \tag{6}$$

- The mean is an "unbiased" estimator, meaning its expected value is equal to the "true value" of the expected value

$$E[\bar{x}] \equiv E[\hat{\mu}] = E[x_i] = \mu \tag{7}$$

- If $x_i \sim N(\mu, \sigma^2)$, the experts tell us that $\bar{x}$ (or$\hat{\mu}$) is Normally distributed $Normal(\mu, \frac{1}{N}\sigma^2)$
- Recall the CLT as a way to generalize this finding: the sampling distribution of the mean is Normal

Estimate the Parameter Sigma

- The Sample Variance is the mean of squared errors

$$sample\ variance(x_i) = \frac{\sum(x_i - \bar{x})^2}{N} \tag{8}$$

- Now the "N-1" problem comes in. This sample variance is not an "unbiased" estimate of σ^2. I mean, sadly,

$$E[sample\ variance(x_i)] \neq \sigma^2 \tag{9}$$

- However, a corrected estimator

$$unbiased\ sample\ variance(x_i) = \frac{\sum(x_i - \bar{x})^2}{N-1} \tag{10}$$

is unbiased:

$$E[unbiased\ sample\ variance(x_i)] = \sigma^2 \tag{11}$$

Standard Error of the Mean

- Two lectures ago, I showed that the variance of the mean is proportional to the true variance of x_i.

$$Var[\hat{\mu}]\, same\ as\ Var[\bar{x}] = \frac{1}{N} Var[x_i] = \frac{1}{N}\sigma^2 \tag{12}$$

(no matter what the distribution of x_i might be).

- We don't know the "true" variance $Var[x_i] = \sigma^2$, but we can take the unbiased sample estimator and use it place of σ^2.
- That gives us the dreaded double hatted estimate of the estimated mean:

$$\widehat{Var[\hat{\mu}]} = \frac{1}{N} unbiased\ sample\ variance(x_i) \tag{13}$$

- You can "plug in" the *unbiased sample variance of x_i* from the previous page if you want to write out a formula!

The magical ratio of $\hat{\mu}$ to $std.err.(\hat{\mu})$

- Because the double hat notation is boring, we call the square root of it the standard error.

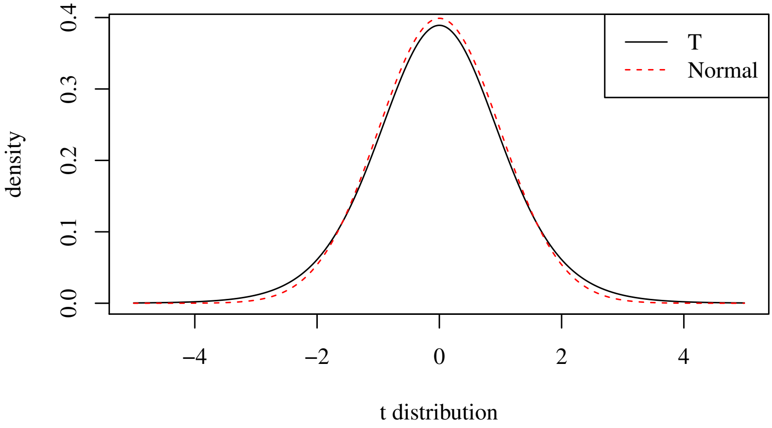
$$std.err.(\bar{x})\ same\ as\ std.err.(\hat{\mu}) = \sqrt{\widehat{Var[\hat{\mu}]}} = \sqrt{\frac{1}{N} unbiased\ sample\ variance(x_i)} \tag{14}$$

- Recall the definition of the term "standard error." It is an estimate of the standard deviation of a sampling distribution.
- Gosset showed that although the true σ^2 is unknown, the ratio of the estimated mean's fluctuations about its true value to the estimated standard deviation of the mean follows a T distribution:

$$\frac{\hat{\mu}-\mu}{\widehat{std.dev.}(\hat{\mu})} = \frac{\hat{\mu}-\mu}{std.err.(\hat{\mu})} \sim T(\nu = N-1) \tag{15}$$

- This new "t variable" becomes our primary interest. Since $Var[x]$ is unknowable, we have to learn to live with the estimate of it, and that brings us down a chain to T.

T distribution with 10 d.f.



T is Similar to Standard Normal, N(0,1)

- symmetric
- single peaked
- But, there is a difference: T depends on a degrees of freedom, $N-1$
  - T is different for every sample size
  - T tends to be “more and more” Normal as the sample size grows

Compare 95% Ranges for Normal and T

```
qnorm(0.025, m=0, s=1)
```

```
[1] -1.959964
```

```
qt(0.025, df=10)
```

```
[1] -2.228139
```

```
qnorm(0.975, m=0, s=1)
```

```
[1] 1.959964
```

```
qt(0.975, df=10)
```

```
[1] 2.228139
```

T-based Confidence Interval

- Using the T distribution, we can "bracket" the 0.95 probability "middle part".
- That puts $\alpha/2$ of the probability outside the 95% range on the left, and $\alpha/2$ on the right
- In a T distribution with 10 degrees of freedom, the range stretches from ($\hat{\mu}$-2.3, $\hat{\mu}$+2.3)
- That's wider than $N(0,1)$ would dictate, of course. The extra width is the penalty we pay for using the estimate $\hat{\sigma}$.

Lets Step through some df values

Note that T is symmetric, so the upper and lower critical points are generally just referred to as $-t_{0.025,df}$ and $t_{0.025,df}$ for a 95% CI with df degrees of freedom

df=20

```
[1] -2.085963  2.085963
```

df=50

```
[1] -2.008559  2.008559
```

df=100

```
[1] -1.983972  1.983972
```

df=250

```
[1] -1.969498  1.969498
```

Summary: The CI for an Estimated Mean Is...

- If
  - $\hat{\mu}$ is Normal, $N(\mu, \sigma^2)$
  - std.err($\hat{\mu}$)= $\hat{\sigma}/\sqrt{N}$ (an estimate of the standard deviation of $\hat{\mu}$)
- Then:

$$CI = [\hat{\mu} - t_{n,\alpha/2} std.err.(\hat{\mu}),\ \hat{\mu} + t_{n,\alpha/2} std.err.(\hat{\mu})] \tag{16}$$

- "*something*" in the CI of the mean is $t_{n,\alpha/2} \times \hat{\sigma}/\sqrt{N}$
- If your sample is over 100 or so, $t_{n,\alpha/2}$ will be very close to 2, hence most of us think of the CI for the mean as

$$[\hat{\mu} - 2\, std.err.(\hat{\mu}),\ \hat{\mu} + 2\, std.err(\hat{\mu})] \tag{17}$$

Symmetric Estimators are easy

- So far as I know, Every estimator that has a symmetrical sampling distribution ends up, one way or another, with a T-based CI.
- Thus, we are preoccupied with finding parameter estimates and standard errors because they lead to CIs that are manageable.
- With NON-symmetric estimators, the whole exercise goes to hell. Everything becomes less generalizable, more estimator-specific, and generally more frustrating ☹.

Outline

Correlation Coefficient

- The product-moment correlation varies from -1 to 1, and 0 means "no relationship".
- The "true" correlation for two random variables is defined as

$$\rho = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{Cov(x,y)}{Std.Dev.(x)Std.Dev.(y)} \tag{18}$$

$$= \frac{E[(x-E[x])\cdot(y-E[y])]}{\sqrt{E[(x-E[x])^2]}\sqrt{E[(y-E[y])^2]}} \tag{19}$$

- Replace those "true values" with sample estimates to calculate $\hat{\rho}$.

How Sample Estimates are Calculated

- Sample Variance: Mean Square of Deviations about the Mean (unbiased version).

$$\widehat{Var[x]} = \frac{\sum_{i=1}^{N}(x_i - \widehat{E[x]})^2}{N-1} \tag{20}$$

- The sample covariance of x and y:

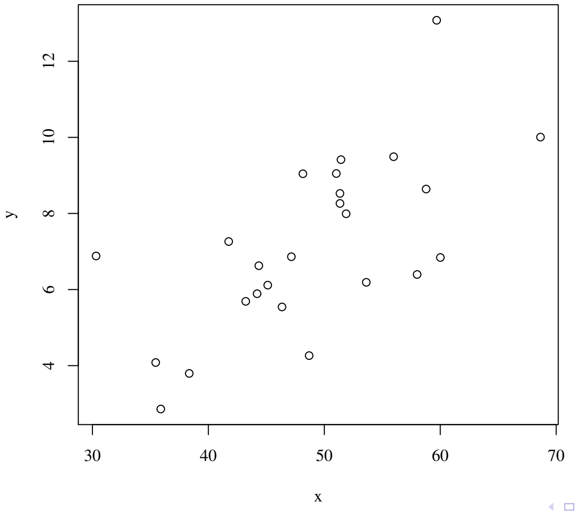
$$\widehat{Cov[x,y]} = \frac{\sum_{i=1}^{N}(x_i - \widehat{E[x]})(y_i - \widehat{E[y]})}{N-1} \tag{21}$$

Covariance: What is that Again?

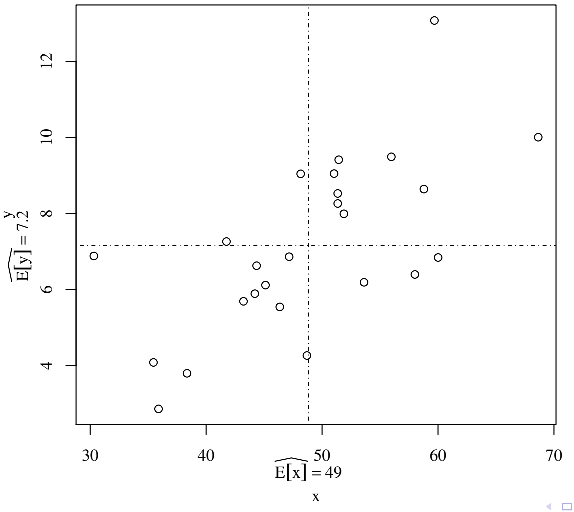
- Intuition:
  - If x and y are both "large", or both "small", then covariance will be positive.
  - If x is "large", but y is "small" (or vice versa), then covariance will be negative.
- The sample "covariance of x with itself" is obviously the same as the variance:

$$\widehat{Cov}[x,x] = \widehat{Var}[x] = \frac{\sum_{i=1}^{N}(x_i - \widehat{E}[x])(x_i - \widehat{E}[x])}{N-1} \tag{22}$$

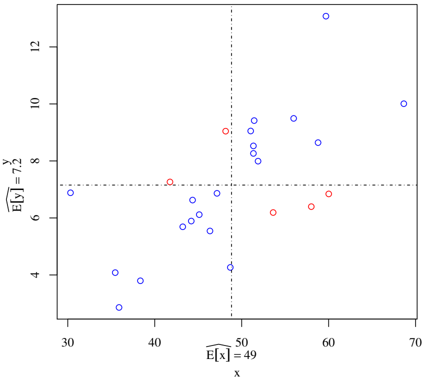
Consider a Scatterplot



Draw in Lines for the Means



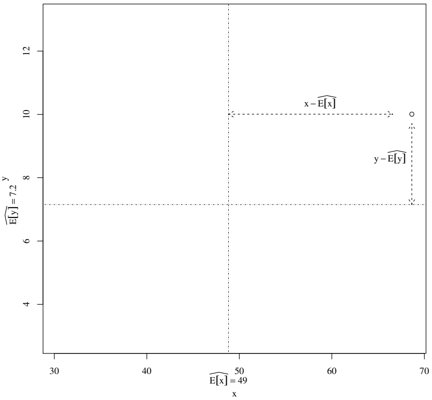
Easier to See Pattern with Some Color



- For each point, necessary to calculate $(x_i - \widehat{E[x]})(y_i - \widehat{E[x]})$
- add those up!
- blue points have positive products
- red points have negative products

+ times + = +, but + times - equals -

- Here, $(x_i - \widehat{E[x]})(y_i - \widehat{E[y]}) > 0$
- Hm. I never noticed before, but that's also the "area" of the rectangle



Remaining Problems

- How do I know 97 is "big" or "medium" number for Covariance
- "How much" will covariance fluctuate from one sample to another, if the parameters of the data generating process remain fixed?

Correlation: Standardize Covariance

- Divide Covariance by the Standard Deviations

$$\frac{\widehat{Cov[x,y]}}{\widehat{Std.Dev.}[x]\cdot\widehat{Std.Dev.}[y]} \tag{23}$$

$$= \frac{\sum(x_i-\widehat{E[x]})(y_i-\widehat{E[y]})/(N-1)}{\left(\sqrt{\sum(x-\widehat{E[x]})^2/(N-1)}\right)\left(\sqrt{\sum(y-\widehat{E[y]})^2/(N-1)}\right)} \tag{24}$$

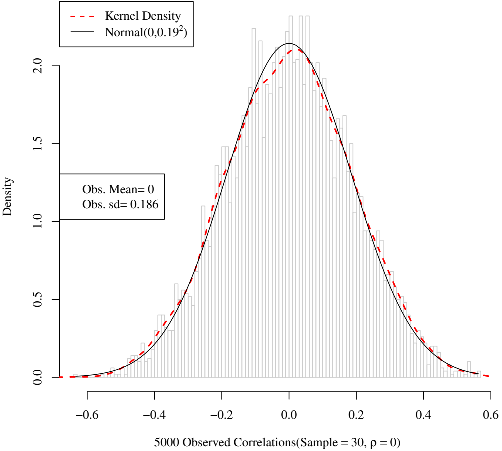
- That produces a number that ranges from -1 to $+1$
  - Check that: Calculate the correlation of x with itself.
- Karl Pearson called it a "product-moment correlation coefficient"
- We often just call it "Pearson's r", or "r".
- Often use variable names in subscript r_{xy} to indicate which variables are correlated.

The Distribution of $\hat{\rho}$ is Symmetric only if ρ is near 0

- If true correlation $\rho = 0$, then the sampling distribution of $\hat{\rho}$ is perfectly symmetric.
- However, if $\rho \neq 0$, the Sampling distribution is not symmetric, and as $\rho \to -1$ or $\rho \to +1$, the Sampling distribution becomes more and more Asymmetric

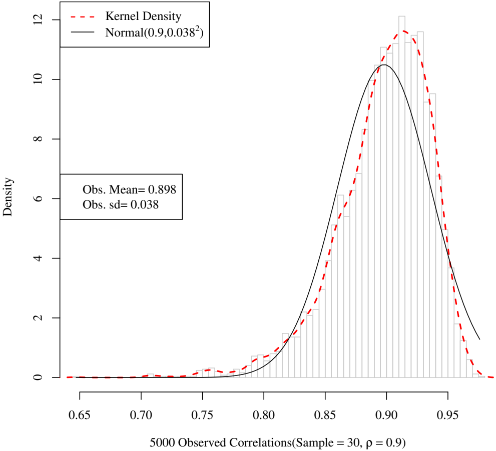
If $\rho = 0$,

- The Sampling Distribution of $\hat{\rho}$ is Symmetric
- Apparently normal, even with small samples.



If $\rho = .90$, $\hat{\rho}$ NOT Symmetric

- The Sampling Distribution of $\hat{\rho}$ is apparently NOT symmetric or normal
- Think for a minute. If the "true rho" is .9, then sampling fluctuation can
  - bump up the observed value only between 0.9 and 1.0
  - bump down the observed value between -1.0 and 0.9

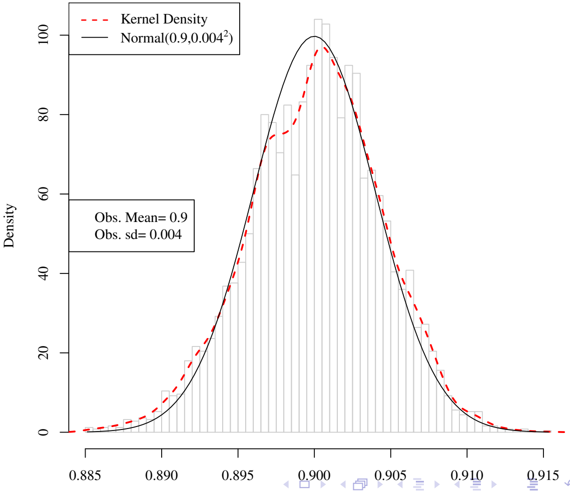


Asymmetric Confidence Interval

- In previous example, the true ρ is 0.9, and the mean of the observed ρ is close to that.
- But the 95% confidence interval is clearly not symmetric.

Can reduce Asymmetry with Gigantic Sample

- Large samples lead to more precise estimates of ρ.
- The sampling distribution of $\hat{\rho}$ is more symmetric when each sample is very large
- Not so non Normal.



Details, Details

- AFAIK, there is no known formula for the exact sampling distribution of $\hat{\rho}$ or its CI
- Formulae have been proposed to get better approximations of the CI
- Fisher proposed this transformation that converts a non-Normal distribution of $\hat{\rho}$ into a more Normal distribution

$$Z = 0.5 ln \left(\frac{1+\hat{\rho}}{1-\hat{\rho}} \right) \tag{25}$$

- The CI can be created in that "transformed space"
- Map back to original scale to get 95% CI.
- Result is an asymmetric CI centered on the sample estimate.

Checkpoint: What's the Point?

- As long as you know the "sampling distribution", you can figure out a confidence interval.
- Work is easier if the CI is symmetric around the estimate $\hat{\theta}$. Usually, with means or regression estimates, the CI is something like

$$\widehat{\theta}\ plus\ or\ minus\ 2 \cdot std.err.(\hat{\theta}) \tag{26}$$

- For Asymmetric sampling distributions, CI have to be approximated numerically (difficult)

Outline

Use π for True Proportion, $\hat{\pi}$ for estimate.

- We already used p for probability and for p-value.
- To avoid conclusion, use π for the Binomial probability of a success
  - π proportion parameter
  - $\hat{\pi}$ a sample estimator
- The "true" probability model is $Binomial(n, \pi)$
- We wish we could estimate π and create a 95% CI

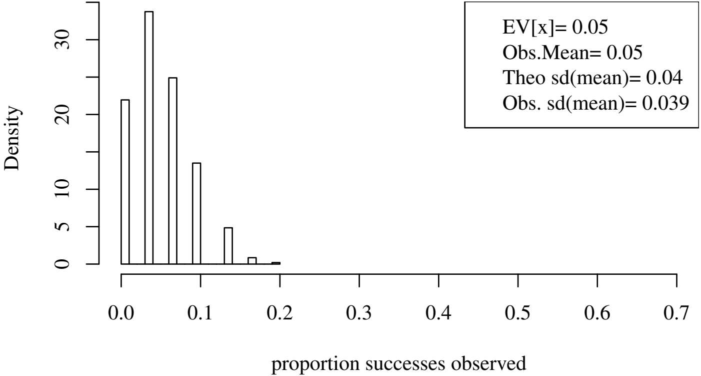
$$\hat{\pi} - something, \; \hat{\pi} + something \tag{27}$$

- But, the sampling distribution is NOT symmetric, so doing that is wrong, which means people who say a CI (margin or error) is *mean plus or minus something* are technically wrong.

Binomial Distribution

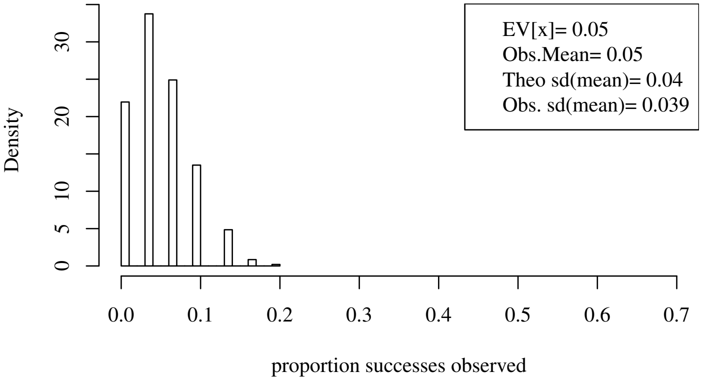
- $Binomal(n, \pi)$ is number of “successes” in n “tests” with probability of success π for each one.
- The observed number of successes from $B(n, \pi)$ is approximately normal if
  - if n is “big enough”
  - and π is not too close to 0 or 1.
- if $\pi = 0.5$, the number of successes $y \sim B(n, \pi)$ is approximately $Normal(n * \pi, \pi(1 - \pi)/n)$,
- The proportion of successes, $x = y/n$, is approximately $Normal(\pi, \pi(1 - \pi))$
- Otherwise, the Binomial is decidedly NOT normal, as we can see from some simulations.

n=30, $\pi = 0.05$; 2000 samples



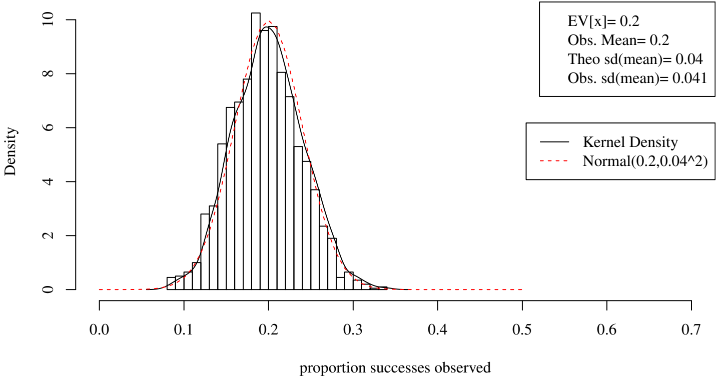
Simulate n=500, $\pi = 0.05$ (2000 estimated proportions)

It doesn't help to make each sample bigger



More Normal with moderate π

Simulate n=100, $\pi = 0.2$ (2000 samples)



Proportions

- The Normal approximation is widely used, but...
- Its valid when N is more than 100 or so and π is in the "mid ranges".
- The Normal approximation lets us take this general idea:

$$CI = [\hat{\pi} - \mathit{something\ low},\ \hat{\pi} + \mathit{something\ high}]$$

- and replace it with

$$CI = [\hat{\pi} - 1.96 \cdot std.error.(\hat{\pi}), \hat{\pi} + 1.96 \cdot std.error(\hat{\pi})]$$

Show My Work: Derive the std.error($\hat{\pi}$)?

This is a Sidenote. Start with the Expected Value

- Recall, for any random variable x,

$$E[x] = \sum prob(x) * x \tag{28}$$

- The chance of a 1 is π and the chance of a 0 is $(1 - \pi)$.
- The expected value of x_i is clearly π:

$$\begin{aligned} E[x] &= \pi * 1 + (1 - \pi) * 0 \\ &= \pi \end{aligned} \tag{29}$$

Show My Work: For the Binomial Case

- The observations are 1's and 0's representing successes and failures: $0, 1, 0, 1, 1, 0, 1$.
- The estimated mean is the "successful" proportion of observed scores

$$\hat{\pi} = \frac{\sum x_i}{N} \tag{30}$$

- Recall this is always true for means, the expected value of the estimate the mean is the expected value of x_i

$$E[\hat{\pi}] = \pi \tag{31}$$

- So it makes sense that we act as though $\hat{\pi}$ is in the center of the CI.

Show My Work: $E[\hat{\pi}] = E[x] = \pi$

This uses the simple fact that expected value is a "linear operator": $E[a \cdot x_1 + bx_2] = aE[x_1] + bE[x_2]$
Begin with the definition of the estimated mean:

$$\hat{\pi} = \frac{x_1}{N} + \frac{x_2}{N} + \ldots + \frac{x_N}{N} \tag{32}$$

$$E[\hat{\pi}] = E\left[\frac{x_1}{N}\right] + \left[\frac{x_2}{N}\right] + \ldots + \left[\frac{x_N}{N}\right] \tag{33}$$

$$E[\hat{\pi}] = N \cdot \frac{E[x]}{N} = E[x] = \pi \tag{34}$$

Show My Work: Variance is Easy Too

- Recall the variance is a probability weighted sum of squared deviations

$$Var[x] = \sum prob(x) * x \tag{35}$$

- For one draw,

$$\begin{aligned} Var[x] = & \quad \pi * (1-\pi)^2 + (1-\pi)(0-\pi)^2 \\ = & \quad (1-\pi)(\pi * (1-\pi) + \pi^2) \\ = & \quad \pi(1-\pi) \end{aligned} \tag{36}$$

- And if we draw N times and calculate $\hat{\pi} = \sum x/N$

$$Var[\hat{\pi}] = \frac{Var[x]}{N} = \frac{\pi(1-\pi)}{N} \tag{37}$$

- Note that's the "true variance", AKA the "theoretical variance" of $\hat{\pi}$.

Show My Work: Here's where we get the standard error

- The standard deviation of $\hat{\pi}$ is the square root of the variance

$$std.dev.(\hat{\pi}) = \sqrt{Var[\hat{\pi}]} = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{N}} \tag{38}$$

- That is the "true standard deviation."
- As we saw in the CLT lecture, the dispersion of the estimator "collapses" rapidly as the sample increases because it is the variance divided by $\sqrt{N}$.
- We don't know π, however. So from the sample, we estimate it by $\bar{x}$ (or, we could call it $\hat{\mu}$).
- Use that estimate in place of the true π and the value is called the standard error

$$std.error(\hat{\pi}) = \sqrt{\pi(1-\pi)}/\sqrt{N}$$

.

Citations on Calculations of CI for Proportions

- These give non-symmetric CI's
  Brown, L. D. Cai, T. T. and DasGupta, A. (2001). "Interval estimation for a binomial proportion." *Statistical Science*, 16(2), 101-133.
  Agresti, A. and Coull, B. A. (1998). "Approximate is better than 'exact' for interval estimation of binomial proportions," *The American Statistician*, 52(2), 119-126.

Outline

What To Remember

- Parameter Estimate, Sampling Distribution, Confidence Interval
- The appeal of the CI is that it gives a "blunt" answer to the question, "how confident are you in that estimate"?
- The symmetric Sampling Distributions usually lead back to the T distribution, which is almost same as $N(0,1)$ for large sample sizes, and a pleasant, symmetric

$$CI = [\hat{\theta} - 2 \cdot std.err.(\hat{\theta})\,,\,,\hat{\theta} + 2 \cdot std.err.(\hat{\theta})] \tag{39}$$

- The nonsymmetric Sampling Distributions do not have symmetric CI's, and the description of their CI's is case specific and contentious.