EMERITUS
INSTITUTE OF MANAGEMENT

WEEK 11

HYPOTHESIS TESTING – ANSWERING QUESTIONS ABOUT YOUR DATA

COLUMBIA | ENGINEERING
EXECUTIVE EDUCATION

Confidence Intervals for Estimating Means

- introduce **confidence intervals** as a way to quantify sampling error
- define and interpret **margin of error**
- how we can use confidence intervals to determine the sample size
  - targeting desired level of precision
- see applications of this to monitoring blood sugar in diabetics

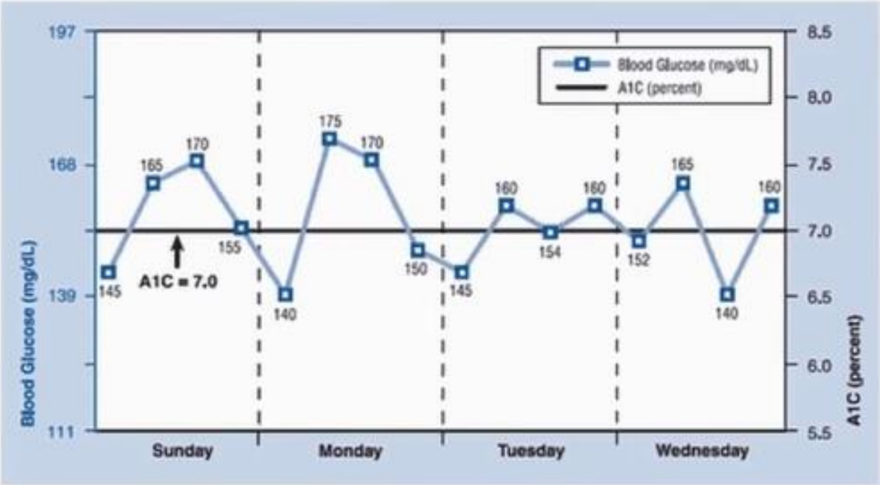
Example: Diabetes

Diabetes (especially type 2) is one of the major epidemics of modern living...

- about 1 in every 10 adults suffers from diabetes (approximately 26M in the US)
- about 90-95% of diabetes is type 2 (which is more easily treatable / preventable)
- significant side effects
- diagnosed diabetes cases cost roughly $300B / annually in the US alone
- there is a genetic component but mostly related to lifestyle (diet, weight, activity)

Monitoring Blood Sugar Levels

- diabetes is diagnosed via blood sugar levels (blood works)
- there are effectively two tests:
  - a localized measurement of blood glucose level
    if $>$ 130 mg/dl (fasting) or $>$ 160 mg/dl (2h after ingestion) then suspect diabetes
  - a time-averaged test based on A1C (if $>$ 6.5 % then suspect diabetes
- the blood glucose level is a measure at a particular point in time
- the A1C test is the **average** glucose level over the past 2-3 months
  - it measures an estimate of the average percent of blood sugar (glucose)
  - 6.5% for A1C is about 140 mg/dl in the standard blood sugar measurement...



Monitoring Blood Sugar Levels

- assume 4 daily measurements
- monitor over two weeks ($n = 56$ observations)
- results:
  - sample mean $\bar{X} = 135$ and standard deviation stdev $= 25$

Q. is this person diabetic?

paraphrase: what is the likelihood that his/her true blood sugar level is above 140?

the 6.5% A1C equivalent threshold...

$\bar{X} = 135$ but it has error

Threshold $= 140$

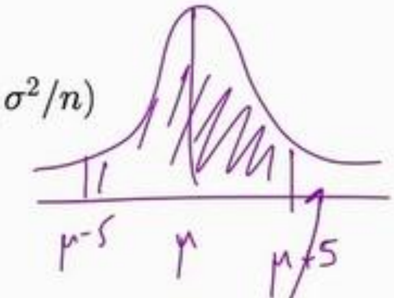
Testing for Diabetes - Formulation

formulation:

- patients **true blood sugar level** μ unknown
  - can only assess using continuous monitoring (not practical...)
- would like to know whether $\mu > 140$.
- have **estimator** of this using **sample mean** equal to 135
  - it is less than 140...

$\boxed{\text{what confidence do we have to rule out diabetes?}}$

Testing for Diabetes - Mechanics



- since we are averaging $n = 56$ observations and we know that $\bar{X} \sim N(\mu, \sigma^2/n)$
  - or Error $= \bar{X} - \mu \sim N(0, \sigma^2/n)$
  - we don't know population σ but we know the sample stdev is 25
- what's the likelihood that $\bar{X}$ is within an error of 5 of the true mean μ?

$$\mathbb{P}\{-5 \leq \bar{X} - \mu \leq 5\} = \mathbb{P}\{-5 \leq \text{Error} \leq 5\}$$

and standardizing

$$\mathbb{P}\left\{\frac{-5}{\sigma/\sqrt{n}} \leq Z \leq \frac{5}{\sigma/\sqrt{n}}\right\} = 1 - 2 \cdot \mathbb{P}\left(Z > \frac{5}{25/\sqrt{56}}\right)$$

with Z being standard normal

$$= 1 - 2 \cdot \mathbb{P}(Z > 1.49)$$

$$= .86$$

- so we can plug in our estimate of σ which is stdev $= 25$ and $n = 56$...
- we get that this likelihood is $P(|\text{error}| \leq 5) = 86\%$

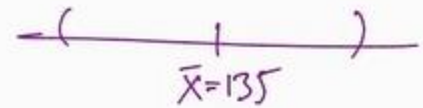
Testing for Diabetes - Mechanics

(mult.)·(Stderror)

- now we turn this around and say that:

"we are 95% confident that the true mean is contained within the interval"

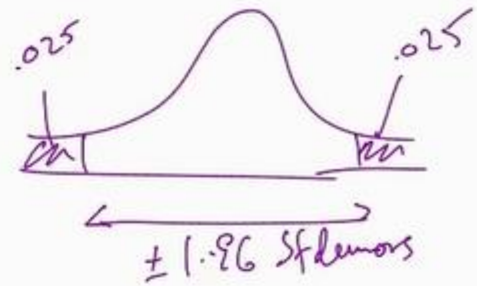
$$\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$



- if we repeat the experiment 100 times, 95 times the true mean will lie in that interval....

- we usually write this as

$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$



the **standard error** is stderror $= (\sigma/\sqrt{n}) = 25/\sqrt{56}$

the **margin of error** is $1.96 \cdot \frac{\sigma}{\sqrt{n}}$

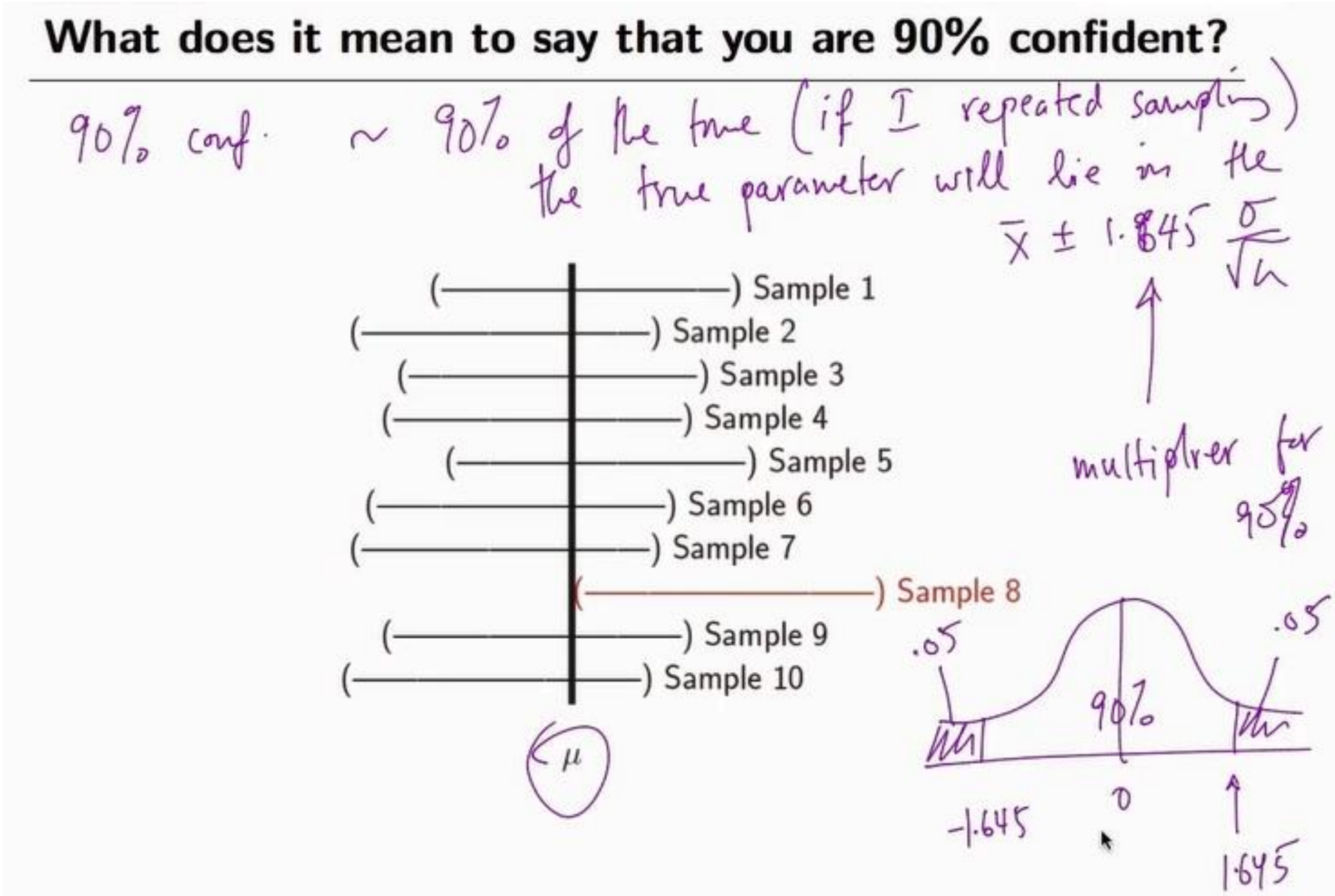
the **confidence level** is 95%

---

rather than fix the margin of error (5 in our example) and find the probability, we fix the confidence level and find the margin of error

eg. 95%

Ninety Percent Confidence



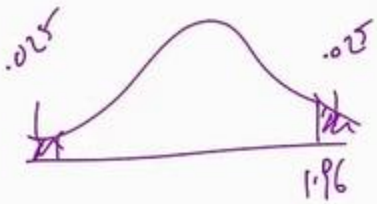
Nine out of ten times, the true parameter would fall within the interval.

Ninety Percent Confidence

$\sigma/\sqrt{n}$

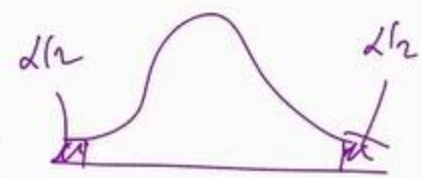
In general,

$$\begin{aligned} \text{CI} &= (\text{point estimate}) \pm (\text{margin of error}) \\ &= \text{point estimate} \pm \text{multiplier} \times \text{Stderror(estimator)} \end{aligned}$$



Confidence level	$\alpha = (1 - \text{CI})$	Multiplier	CI
90%	.10	1.645	
95%	.05	1.96	$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
99%	.01	2.58	

notation: let z_w be the point such that $\mathbf{P}(Z \geq z_w) = w$
i.e., area to the right of z_w is w



In general CIs are expressed in terms of α:
for a level-α CI, multiplier $= z_{\alpha/2}$

$z_{\alpha/2}$

Recipe: CI for the Population Mean μ

setup: sample $X_1, \ldots, X_n$, taken from a population with mean μ and variance σ^2

1. compute estimator (sample mean) $\bar{X} = (X_1 + \cdots + X_n)/n$

2. choose α (i.e., confidence level to achieve)

3. find $z_{\alpha/2}$

4. the $(1 - \alpha)$ CI if we know σ is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

5. in practice, use sample standard deviation (stdev) in place of (typically) unknown σ

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

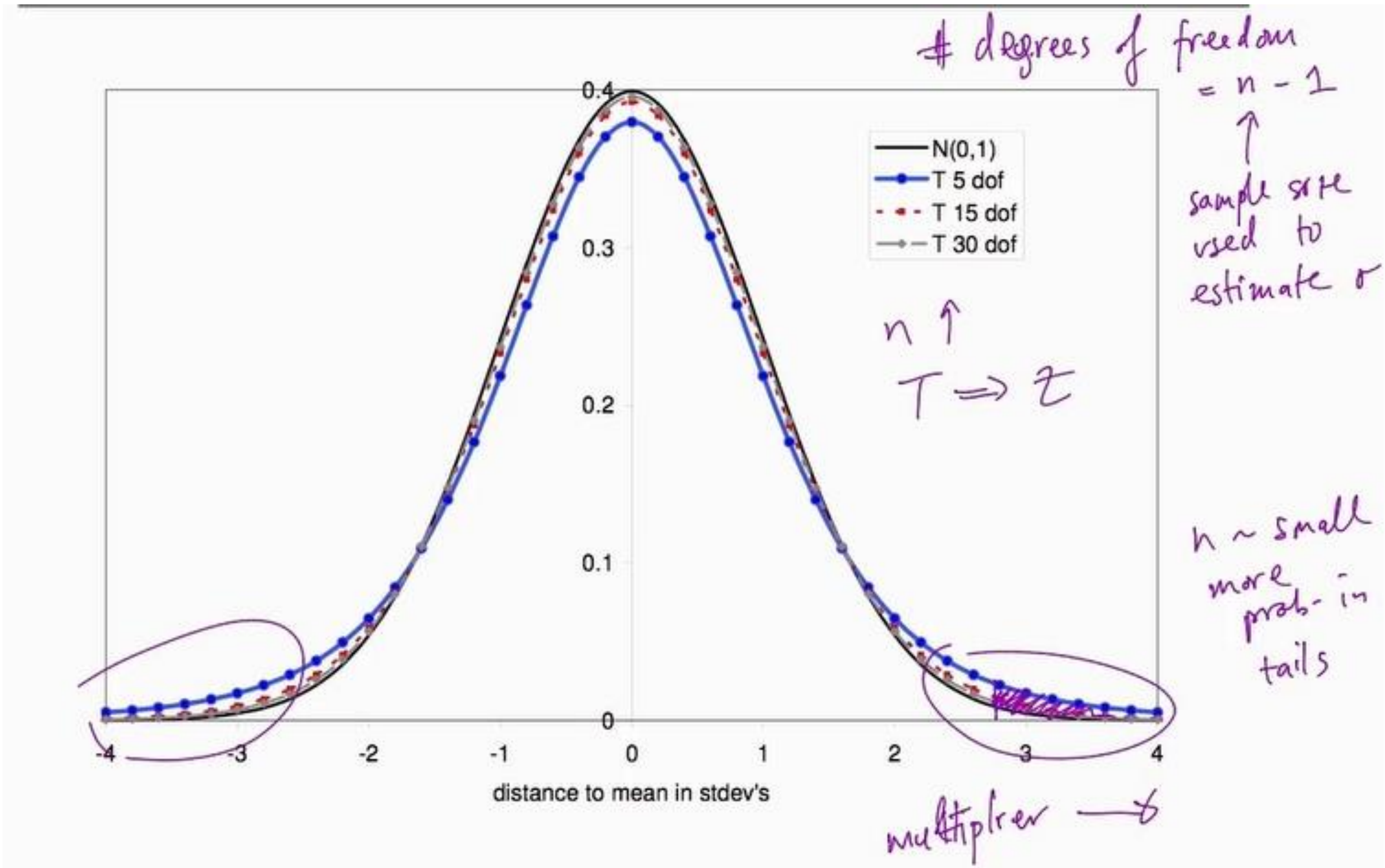
   $s =$ STDEV(...) in EXCEL

T-Tables and T-Multipliers

Issue: when we replaced the true (and unknown) standard deviation σ (for the full population) with the sample standard deviation s = STDEV(...) from our sample, we introduce an additional error

- we can compensate for this by using a larger multiplier in our confidence intervals
- instead of the Z-value, from the normal table, we use the value from the t distribution
  - for sample size n we use $n - 1$ degrees of freedom
- general rules of thumb:
  - can ignore this correction if sample size is at least 30 [see table], **but...**
  - ...in regression we will always use t-tables
  - correction **not valid** if original data is discrete/ordinal (like polls)

Comparing Normal and Distribution and T-Distribution



Interpretation of CIs

$$\text{CI} = \text{point estimate} \pm \text{multiplier} \times \text{Stderror}$$

what determines the margin of error?

1. Multiplier depends on conf. level. conf. Level ↑, mult. ↑
2. Sample size: $\text{Stderr} = \sigma/\sqrt{n}$ as n ↑, stderr ↓ like $\frac{1}{\sqrt{n}}$
3. Variability of underlying data (σ) in the stderr $= \sigma/\sqrt{n}$

How do we choose the sample size n in order to tighten our $(1-\alpha)$ CI?

$$(\text{margin error}) = \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha/2}$$

$$(\text{Stderr}) \cdot (\text{Mult}).$$

Sample Size Determination

- the A1C test reports results that are with 95% confidence $\pm 0.5\%$ (about ± 10 mg/dl)
  - so if you receive a result of 6.5% the actual A1C may be 6%... below threshold

Q. how many samples n of blood glucose level do we need to take to get a margin of error of ± 10 (at 95% confidence), which would correspond to the accuracy of the A1C measurement?

- $1.96 \cdot 25/\sqrt(n) = 10$ [10 mg/dl is the std error in the A1C test]
- solving for n gives $n = (1.96 \cdot 25/10)^2 = 25$ (we had 56 in our sample...)
- more generally, at 95% confidence

$$\textbf{required sample size } n = \left(\frac{1.96 \cdot \sigma}{\text{Margin of Error}}\right)^2$$

- squaring means the required sample size grows quickly if we want very precise results
- we also need an estimate for σ (stdev) or small pilot study...

Summary: Confidence Intervals

We would like to complement
our **point estimate** $\bar{X}$ or $\hat{p}$ with **an interval**

"We are 95% confident that the true parameter (μ or p) lies
in a certain interval"

Method:

- all CIs are computed in the same way:

$$\begin{aligned} \text{CI} &= \text{point estimate} \pm \text{margin of error} \\ &= \text{point estimate} \pm \text{multiplier} \times \text{Stderror} \end{aligned}$$

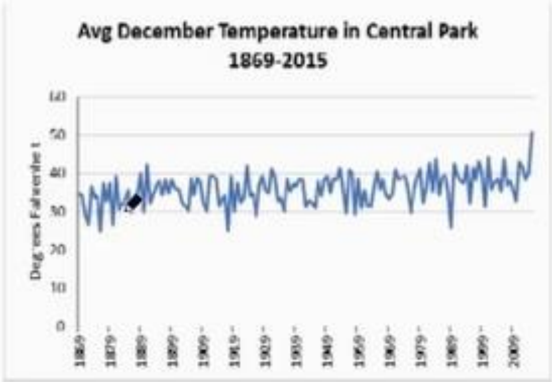
- Stderror is the Stdev of the estimator ($\bar{X}$ or $\hat{p}$)

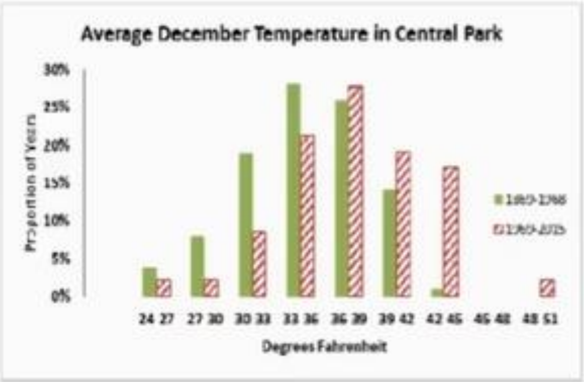
Summary: Confidence Levels

- confidence intervals for difference in means and proportions
- statistical significance
- p-value
- A/B testing

- **statistical significance** measures "strength of statistical evidence" in support of some claim
- p-value as a measure of **statistical significance**
- the smaller the p-value the stronger the statistical significance of the evidence
- the opposite of "statistically significant" is "due to chance" (spurious/ fluke)
- statistically significant does not (necessarily) mean "important"

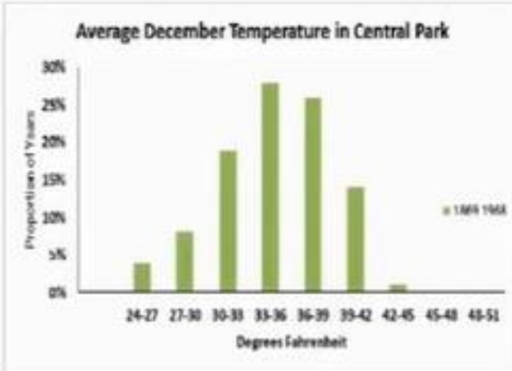
Central Park: Weather Data

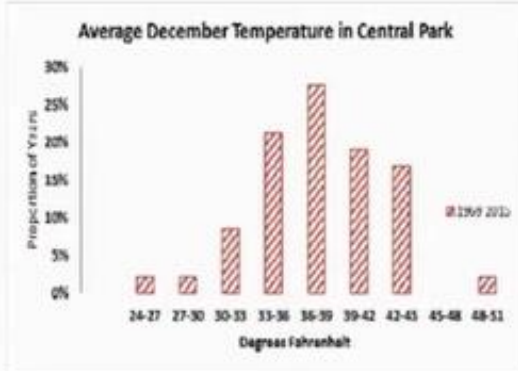




- is there an upward trend?
- has the mean shifted over time?
- is the apparent increase in mean statistically significant given the high degree of variability?

Central Park: Weather Data



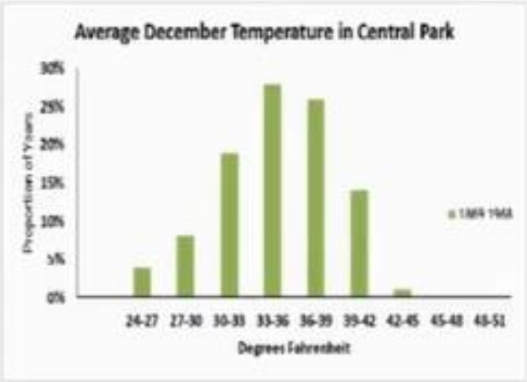


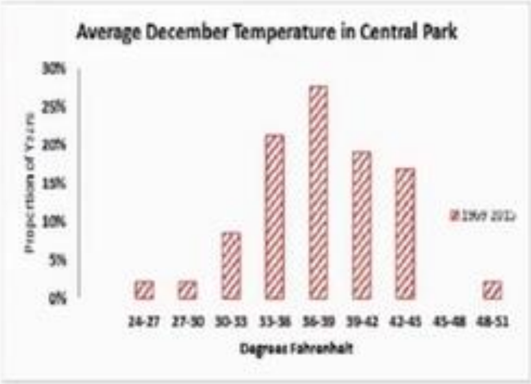
	1869-1968	1969-2015	1969-2014	Full History
Mean	35.0	38.1	37.8	36.0
Median	34.8	38.4	38.4	35.9
Stdev	3.8	4.4	4.0	4.3

- is the observed increase in average temperature statistically significant?
- paraphrasing: is the difference in means "large" relative to the variability in the data?

Central Park: Weather Data

Differentiation in mean





	1869-1968	1969-2015	1969-2014	Full History
Mean	35.0	38.1	37.8	36.0
Median	34.8	38.4	38.4	35.9
Stdev	3.8	4.4	4.0	4.3

- $Y_1, \ldots, Y_{100}$ are the observations (years) for the earlier data [n=100]
- $X_1, \ldots, X_{47}$ are the observations for the more recent data [m=47]

the difference in means is: $\bar{X} - \bar{Y} = 38.1 - 35.0 = 3.1$

how do we construct a confidence interval for this?

Central Park: Weather Data

- the standard error for the *difference in means*

$$\text{stderror}[\bar{X} - \bar{Y}] = \sqrt{\text{stderror}[\bar{X}]^2 + \text{stderror}[\bar{Y}]^2} = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

$Var(X-Y) = Var X + Var Y$

Use this formula

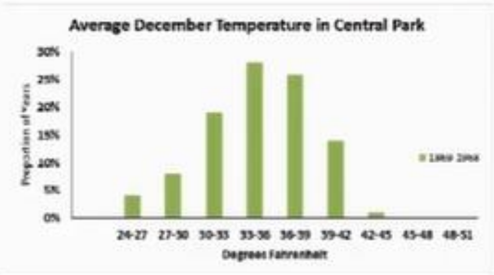
- the 95% confidence interval is:

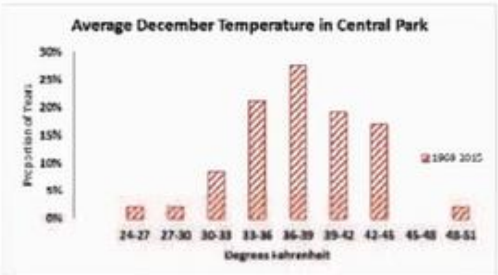
$$(\bar{X} - \bar{Y}) \pm 1.96 \cdot \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

3.1

  - s_X = STDEV($X_1, \ldots, X_n$), the recent years sample standard deviation
  - s_Y = STDEV($Y_1, \ldots, Y_m$), the earlier years sample standard deviation

Confidence Levels to P-Values





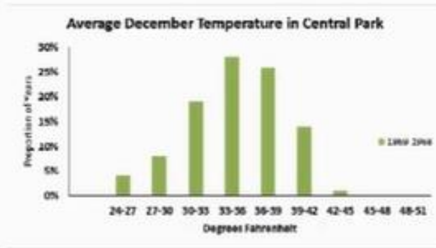
	1869-1968	1969-2015	Difference	
Mean	35.0	38.1	3.1	
Stdev	3.8	4.4		
No. Years	100	47		
Std Err	0.38	0.64	0.75	<--SAXBY

$$\sqrt{\frac{3.8^2}{100}+\frac{4.4^2}{47}}$$

confidence interval for difference: $3.1 \pm 1.96 \cdot 0.75 = 3.1 \pm 1.5$

$= (1.6, 4.6)$

Confidence Levels to P-Values



	1869-1968	1969-2015		Difference	
Mean	35.0	38.1		3.1	
Stdev	3.8	4.4			
No. Years	100	47			
Std Err	0.38	0.64		0.75	<--SAXBY

$$\sqrt{\frac{3.8^2}{100} + \frac{4.4^2}{47}}$$

confidence interval for difference: $3.1 \pm 1.96 \cdot 0.75 = 3.1 \pm 1.5$

$= (1.6, 4.6)$

Q. what happens if the confidence interval would straddle zero?

A. in that case the **true** difference in means could be zero

$\boxed{\text{we can't tell the two means apart!}}$

- in that case the evidence that Central Park is getting warmer is not statistically significant

Confidence Levels to P-Values

Q. when will this happen in our case?

- what if we up the confidence level to 98%?

  the z-multiplier will be 2.33 and the CI will be $3.1 \pm 2.33 \cdot 0.75 = 3.1 \pm 1.75$

- what if we up the confidence level to 99%?

  the z-multiplier will be 2.57 and the CI will be $3.1 \pm 2.57 \cdot 0.75 = 3.1 \pm 1.93$

- what if we up the confidence level to 99.99%?

  the z-multiplier will be 3.27 and the CI will be $3.1 \pm 3.27 \cdot 0.75 = 3.1 \pm 2.45$

- what if we up the confidence level to 99.997%? [it's not even in your z-tables...]

  the z-multiplier will be 4.2 and the CI will be $3.1 \pm 4.2 \cdot 0.75 = 3.1 \pm 3.15$

  finally the CI straddles zero!

- we report this as a p-**value** of 0.003%

Confidence Levels to P-Values

Q. when will this happen in our case?

$\frac{3.1}{.75} \sim 4.15$ Stdem

- what if we up the confidence level to 98%?

  the z-multiplier will be 2.33 and the CI will be $3.1 \pm 2.33 \cdot 0.75 = 3.1 \pm 1.75$

- what if we up the confidence level to 99%?

  the z-multiplier will be 2.57 and the CI will be $3.1 \pm 2.57 \cdot 0.75 = 3.1 \pm 1.93$

- what if we up the confidence level to 99.99%?

  the z-multiplier will be 3.27 and the CI will be $3.1 \pm 3.27 \cdot 0.75 = 3.1 \pm 2.45$

- what if we up the confidence level to 99.997%? [it's not even in your z-tables...]

  the z-multiplier will be 4.2 and the CI will be $3.1 \pm \underline{4.2} \cdot 0.75 = \underline{3.1} \pm \underline{3.15}$

  finally the CI straddles zero!

At 99.997% level we can no longer conclude that $\mu_x - \mu_y > 0$

- we report this as a p-**value** of 0.003%

Confidence Levels to P-Values

99.997%

p-value (in %): is 100 - (level of confidence) where the confidence interval straddles zero

interpretation of p-value:

- smaller p-value means more statistically significant
  - usually the threshold for saying something is "statistically significant" is p-value of 0.05
  - anything below 0.05 means statistically significant
- p-value is the probability that the difference we see in sample means is due to chance
  - hence the smaller the p-value the less likely the difference is a fluke...

Likelihood that the sample mean diff. = + 3.1 degr. where $(\mu_x \leq \mu_y)$

Q. what about the central park data?

our p-value says that the likelihood the 3.1 degree increase in recent years is due to random chance (i.e., Central Park isn't getting warmer) is 0.003%

- we conclude that it's **extremely unlikely** that Central Park isn't getting warmer... $= .003\%$

Computing P-Values: Simpler Way

- playing around with the confidence dial is a cumbersome way to compute the p-value...
- finding the point where the confidence interval straddles zero is equivalent to

$$\bar{X} - \bar{Y} = (\text{z-value}) \cdot \text{stderror}[\bar{X} - \bar{Y}]$$

- we solve this for the z-value and call it the **test statistic** or **t-stat**

$$[(\text{z-value}) =] \quad \textbf{t-stat} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

- then we see how much area lies in the two tails of the normal table and this is the p-value

$$p\text{-value} = 2\mathbb{P}\{Z \geq |\text{t-stat}|\}$$

- in the Central Park data we had a difference of means of 3.1 and stderror of 0.75 so:
  - z-value $= 3.1/0.75 = 4.133$
  - p-value $= 2\mathbb{P}\{Z \geq 4.133\} = 0.00003$ [or 0.003%]

Impact of Online Ad Analysis

- Click-through rate (CTR) of the Citibank ad?
- $\text{CTR} = \dfrac{\text{number of clicks}}{\text{number of impressions}} = 0.01\%$

- $\text{CTR} = 0.05\%$
- Clearly 0.01% $<$ 0.05%, but is this a "systematic" difference in proportions, or a difference likely due to chance? But first of all, can we even compare these numbers?

Correlation and Causation

- Correlation is not causation!
  - Think of unobserved variables that can be confounding the effect of the ad
- Above, we are not comparing "apples to apples"
- What can we do about it?
  - Run a randomized experiment!

Why Experiments

- Random assignment of subjects to treatment and control guarantees that the treatment and control groups are comparable in *every way* except in the reception of the treatment

- As a result, we can safely attribute differences in the outcomes to differences in the treatment as opposed to differences in other unobserved factors

- In simple words, the flip of a fair coin knows nothing about the characteristics of a subject, so it tends to be equitable: it tends to produce treatment and control groups that are similar

- For this reason, randomization is the cornerstone of modern experimentation with human subjects
  - Think about clinical trials
  - In the Internet settings, think about A/B tests

A/B Testing

- A/B testing has been referred to as a fundamental change in strategy for business decision-making
  - A turn towards evidence-based decision-making
  - For example, at Facebook data scientists run over 1000 experiments each day

- What has driven this change?
  - On the Internet, small improvements can translate into massive profits given its large scale
  - Running A/B tests is cheap

- A/B testing is a term for a randomized experiment with two "treatments" or variants
  - A "bake-off" between competing variants
  - A/B tests can be extended to three or more variants

E-Mail Campaign Efficacy

- Want to email customer base to increase sales through its webpage
- Script two emails —identical in every way— except in the following wording:
  - Email 1: "Limited time offer! Use promo code: `ABC123`"
  - Email 2: "Offer expires on Sunday! Use promo code: `123ABC`"
- Send each of the emails to 50,000 different recipients and measure response.
  - Email 1: 1% visit rate; 0.05% buy rate
  - Email 2: 0.5% visit rate; 0.03% buy rate
- Questions:
  - 1% > 0.5%, but is this difference *statistically significant*? In other words, is this a systematic difference, not due to random chance?
  - Is the difference between the buy rates statistically significant?
  - What sample size would allow to detect differences of size 0.02% with 95% confidence?

Confidence Intervals for Difference in Proportions

- Remember the basic structure of a confidence interval

$$\begin{aligned}\text{confidence interval} &= \text{point estimate} \pm \text{margin of error}\\ &= \text{point estimate} \pm \text{multiplier} \times \text{stderror[estimator]}\end{aligned}$$

$$\sqrt{\frac{.01(1-.01)}{50{,}000} + \frac{.005(1-.005)}{50000}}$$

- Here
  - point estimate $= \hat{p}_1 - \hat{p}_2$
  - multiplier $= 1.96$ (for 95% confidence)
  - stderror[estimator] $= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

- In this way, the confidence intervals for difference in two proportions is

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

1% .5%

Statistical Difference: Rate of Visit

- Results:
  - Email 1: $n_1 = 50{,}000$; 1% visit rate; 0.05% buy rate
  - Email 2: $n_2 = 50{,}000$; .5% visit rate; 0.03% buy rate

$$(.01 - .005) \pm 1.96 \sqrt{\frac{.01(1-.01)}{50000} + \frac{.005(1-.005)}{50000}}$$

$$\approx .005 \pm 1.96 \sqrt{\frac{.015}{50000}} \sim .5\% \pm .1\%$$

95% Conf. that the CTR with email 1 is .4% & .6% higher?

Statistical Difference: Rate of Visit

Interpretation

- The confidence interval for the difference in visit rates is 0.5% $\pm$ 0.11% or [0.39%, 0.61%]
- This implies the difference in *population* proportions $p_1 - p_2$ is contained in the interval [0.39%, 0.61%] with 95% confidence
- The confidence interval does not contain zero: p_1 is greater than p_2 by at least 0.39 percentage points

> In this case we say that "the difference between the two proportions is **statistically significant** at the 5% level"

- The meaning of this statement is:

"There is only a 5% chance that the difference of 0.5 percentage points is caused by chance, and there is 95% likelihood the two population proportions are different."

- Conclusion: visit rates are significantly higher with Email 1 than with Email 2 at the 5% level

Statistical Difference: Buying

- Results:
  - Email 1: $n_1 = 50{,}000$; 1% visit rate; 0.05% buy rate
  - Email 2: $n_2 = 50{,}000$; 0.5% visit rate; 0.03% buy rate

Interpretation

- 95% CI for the difference in buy rates is 0.02% $\pm$ 0.025% or [-0.005%, 0.045%]

- . . . the difference in population proportions $p_1 - p_2$ is contained in the interval [-0.005%, 0.045%] with 95% confidence

- Now the confidence interval contains zero!

> In this case, we say that "the difference between the two proportions is **not** statistically significant at the 5% level"

- In simple words, the **true** difference in proportions could be zero. We can't tell the two proportions apart!

- Conclusion: the buy rates for the two emails are not significantly different at the 5% level

Market Segmentation

- Now consider the following results by gender

	Gender		
	Men	Women	Total
Email 1	0.47%	0.53%	1.00%
Email 2	0.24%	0.27%	0.50%

- And now by gender and age group

Email 1				Email 2			
	Gender				Gender		
Age group	Men	Women	Total	Age group	Men	Women	Total
18-24	0.028%	0.032%	0.060%	18-24	0.012%	0.013%	0.025%
25-34	0.056%	0.064%	0.120%	25-34	0.026%	0.029%	0.055%
35-44	0.080%	0.090%	0.170%	35-44	0.042%	0.048%	0.090%
45-54	0.103%	0.117%	0.220%	45-54	0.054%	0.061%	0.115%
55-64	0.089%	0.101%	0.190%	55-64	0.042%	0.048%	0.090%
65-74	0.066%	0.074%	0.140%	65-74	0.033%	0.037%	0.070%
75 or older	0.047%	0.053%	0.100%	75 or older	0.026%	0.029%	0.055%

Sample Size of Segments

- Sample sizes by gender and age group

Email 1

Age group	Men (Gender)	Women (Gender)	Total
18-24	1,410	1,590	3,000
25-34	2,820	3,180	6,000
35-44	3,995	4,505	8,500
45-54	5,170	5,830	11,000
55-64	4,465	5,035	9,500
65-74	3,290	3,710	7,000
75 or older	2,350	2,650	5,000
Total	23,500	26,500	50,000

Email 2

Age group	Men (Gender)	Women (Gender)	Total
18-24	1,175	1,325	2,500
25-34	2,585	2,915	5,500
35-44	4,230	4,770	9,000
45-54	5,405	6,095	11,500
55-64	4,230	4,770	9,000
65-74	3,290	3,710	7,000
75 or older	2,585	2,915	5,500
Total	23,500	26,500	50,000

Not sufficient data size

- Imagine segmenting on more variables such as city, race, web browsing history...
- Clearly, for targeting very specific segments we need very large data sets
- Or to rely on a model (e.g., linear regression)

Key Takeaways

- Confidence interval for a difference in means:

$$\bar{X} - \bar{Y} \pm 1.96 \times \sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}$$

- Confidence interval for a difference in proportions:

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Statistical significance measures the "strength of statistical evidence" in support of some claim

- The p-value is a measure of statistical significance
  - The p-value is the smallest value of α such that the confidence interval does not include 0 or another hypothesized value
  - The smaller the p-value, the stronger the evidence that our estimate is different to the hypothesized value

www.emeritus.org