Univariate Data Description: One Variable at a Time

- Topics:
  - Summarizing categorical and quantitative variables
    - Calculating mean, variance, and skewness statistics
    - Other summary measures for skewed quantitative variables
  - Sampling distributions for sample statistics:
    - Quantifying uncertainty in sample means
    - Inferences from sample means to expected population means
    - Bonus: sampling distributions of variances

Univariate Descriptors by Type of Variable

- For now we focus on the possible values of each variable given how it was measured, and thus by what salient features we should describe it **univariately** (“uni” = one by itself)
  - Two main types of variables: categorical or quantitative

- **Categorical** (numbers are labels): Binary, Ordinal, or Nominal
  - Just need to know **frequency** of each category
  - Often reported as **percent**: frequency divided by total possible
  - Can be displayed graphically using a **frequency plot** (bar graph)
  - **Value labels** make this information easier to digest or present

Example Variable for Marital Status: Request Frequencies and Percentages

In **SAS**, using **PROC FREQ**:

```
PROC FREQ DATA=work.Example1;
TABLE marital;
RUN;
```

Marital: 5-Category Marital Status				
marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.Married	900	**45.59**	900	45.59
2. Widowed	163	8.26	1063	53.85
3. Divorced	317	16.06	1380	69.91
4. Separated	68	3.44	1448	73.35
5. Never Married	526	26.65	1974	100.00

In **STATA**, using **TABULATE**:

```
tabulate marital

marital: 5-Category |
     Marital Status |      Freq.     Percent        Cum.
--------------------+-----------------------------------
          1.Married |        900       45.59       45.59
          2.Widowed |        163        8.26       53.85
         3.Divorced |        317       16.06       69.91
        4.Separated |         68        3.44       73.35
    5.Never Married |        526       26.65      100.00
--------------------+-----------------------------------
              Total |      1,974      100.00
```

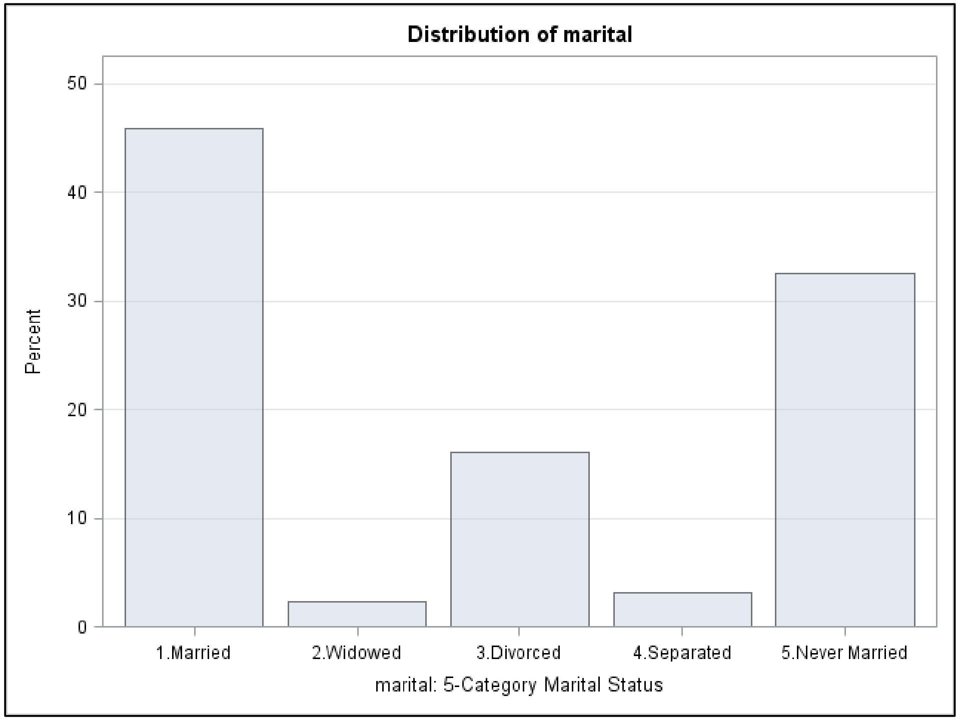
Note that in HW 1 and 2, these **percentages will need to be entered as proportions out of 1**. For instance, 45.59% should be entered as 0.4559 instead of 45.59.

Example Variable for Marital Status: Request a Frequency Plot (Bar Graph)

- **In SAS:**

```
PROC FREQ DATA=work.Example1;
TABLE marital / PLOTS=FREQPLOT(TYPE=BAR SCALE=PERCENT);
RUN;
```

  - x-axis (horizontal) shows each observed category
  - y-axis (vertical) shows percentage for each category
  - Value labels provide meaning of numbers



- Also in **STATA**, using **HISTOGRAM**

```
histogram marital, discrete percent xla(1/5, valuelabel alternate)
```

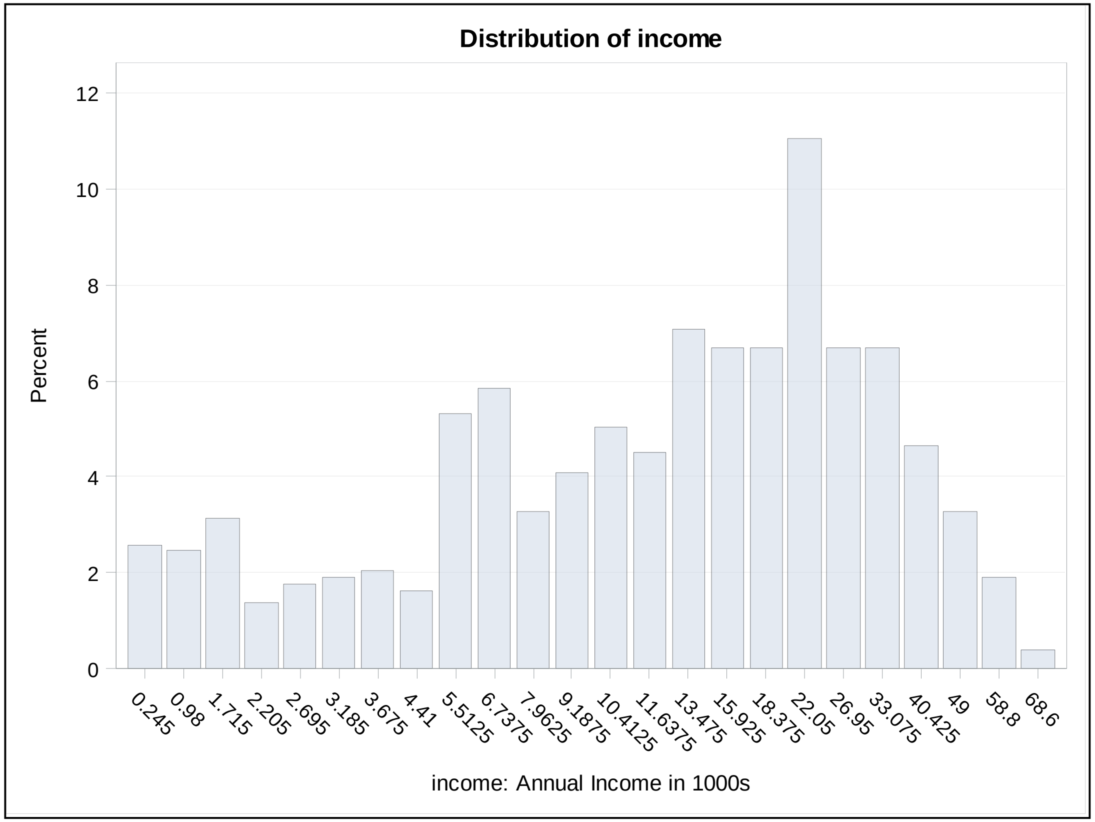
What about Quantitative Variables?

- **Quantitative variable**: numbers are numbers! (interval measurement)
  - May be bounded (binomial, count) or "continu-ish"
- For quantitative variables with **many observed values**, a frequency list of each distinct value is less useful (interval is ignored)
  - For instance, consider annual income in $1000s (from multiple choices, so it's "continu-ish" here):

income: Annual Income in 1000s				
income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0.245	19	2.59	19	2.59
0.98	18	2.45	37	5.04
1.715	23	3.13	60	8.17
2.205	10	1.36	70	9.54
2.695	13	1.77	83	11.31
3.185	14	1.91	97	13.22
3.675	15	2.04	112	15.26
4.41	12	1.63	124	16.89
5.5125	39	5.31	163	22.21
6.7375	43	5.86	206	28.07
7.9625	24	3.27	230	31.34
9.1875	30	4.09	260	35.42
10.4125	37	5.04	297	40.46
11.6375	33	4.50	330	44.96
13.475	52	7.08	382	52.04
15.925	49	6.68	431	58.72
18.375	49	6.68	480	65.40
22.05	81	11.04	561	76.43
26.95	49	6.68	610	83.11
33.075	49	6.68	659	89.78
40.425	34	4.63	693	94.41
49	24	3.27	717	97.68
58.8	14	1.91	731	99.59
68.6	3	0.41	734	100.00

What about Quantitative Variables?

- Frequency plot: also not helpful…



The values are being treated as distinct categories without regard to the intervals between them…

income: Annual Income in 1000s				
income	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0.245	19	2.59	19	2.59
0.98	18	2.45	37	5.04
1.715	23	3.13	60	8.17
2.205	10	1.36	70	9.54
2.695	13	1.77	83	11.31
3.185	14	1.91	97	13.22
3.675	15	2.04	112	15.26
4.41	12	1.63	124	16.89
5.5125	39	5.31	163	22.21
6.7375	43	5.86	206	28.07
7.9625	24	3.27	230	31.34
9.1875	30	4.09	260	35.42
10.4125	37	5.04	297	40.46
11.6375	33	4.50	330	44.96
13.475	52	7.08	382	52.04
15.925	49	6.68	431	58.72
18.375	49	6.68	480	65.40
22.05	81	11.04	561	76.43
26.95	49	6.68	610	83.11
33.075	49	6.68	659	89.78
40.425	34	4.63	693	94.41
49	24	3.27	717	97.68
58.8	14	1.91	731	99.59
68.6	3	0.41	734	100.00

What about Quantitative Variables?

- **Instead we need a histogram**, which combines observations on the x-axis into “bins” (that you can and should choose!)
  - For example: income in $1000s in **bins from 0 to 70 in increments of 5**

- In **SAS**:

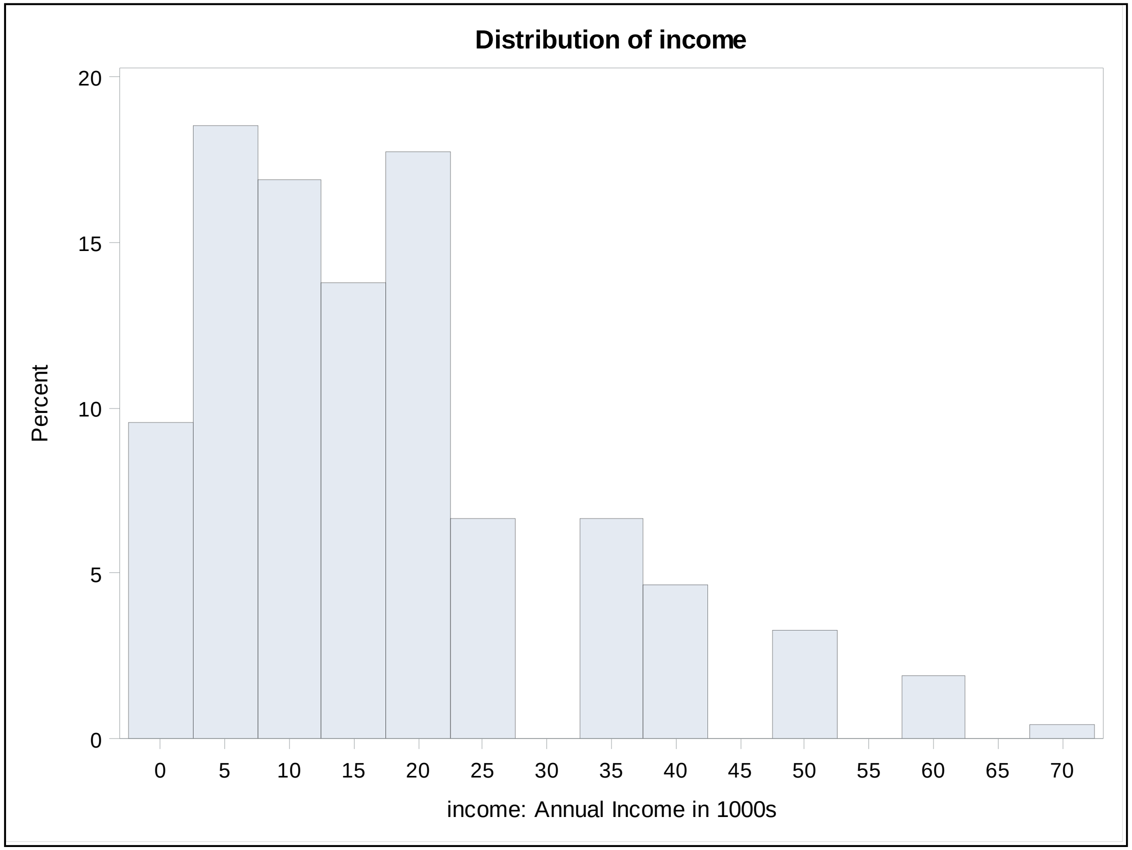
```
PROC UNIVARIATE DATA=work.Example1;
VAR income; * VAR means variable;
HISTOGRAM income / MIDPOINTS=0 TO 70 BY 5;
RUN;
```

- In **STATA**: `histogram income, percent discrete width(5) start(0)`

- Not as easy to make histograms in Excel (have to combine observations into bins manually first, then make bar chart)

What about Quantitative Variables?

- **Instead we need a histogram**, which combines observations on the x-axis into "bins" (that you can and should choose!)
  - For example: income in $1000s in **bins from 0 to 70 in increments of 5**
  - Number and width of bins will be chosen for you otherwise
  - x-axis (horizontal) shows bins of values
  - y-axis (vertical) shows percentage within each bin



Quantitative Variables: 3 Salient Summary Features

1. **Central tendency**: think "middle of distribution"; can be given by:
   - Mean = arithmetic average (abbreviated "M" in results)
   - Also by Median = middle value if ordered from most to least
   - Also by Mode = most frequent value

2. **Dispersion**: think "width of distribution", can be given by:
   - Standard Deviation (abbreviated "SD" in results) = average deviation of any given observation (e.g., person) from the mean
   - Variance (abbreviated "VAR" in results) = *squared* average deviation of any given observation (e.g., person) from the mean (so $VAR = SD^2$)
   - Also by Inter-Quartile Range = distance from 25th to 75th percentile

3. **Skewness** = asymmetry (more values on one side than the other)
   - Is often caused by natural boundaries in practice (e.g., counts at 0)
   - Is something to factor into your analysis, but is not usually reported

Calculating the Arithmetic* Mean of Quantitative (or Binary) Variables

- New notation:
  - $\boldsymbol{y_i}$= "y sub i" = outcome y for person i
  - $\boldsymbol{N}$ = "big N" = number of persons in the sample
  - $\boldsymbol{y_N}$= "y sub N" = last person in the sample
  - $\boldsymbol{\bar{y}}$ = "y bar" = sample arithmetic* mean
    - Note the lack of an i subscript—this is because $\bar{y}$ is a constant, not a variable

- Using new notation, how to calculate **sample mean** (M in results):

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$$

→ "Start at $i = 1$, sum over all the y values ending at N, then divide that total by N"

** Yes, there are other kinds of means (geometric, harmonic, weighted)…*

Calculating the Variance of Quantitative Variables

- Using notation to calculate the **variance** (VAR in results):

$$Variance = s^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N-1}$$

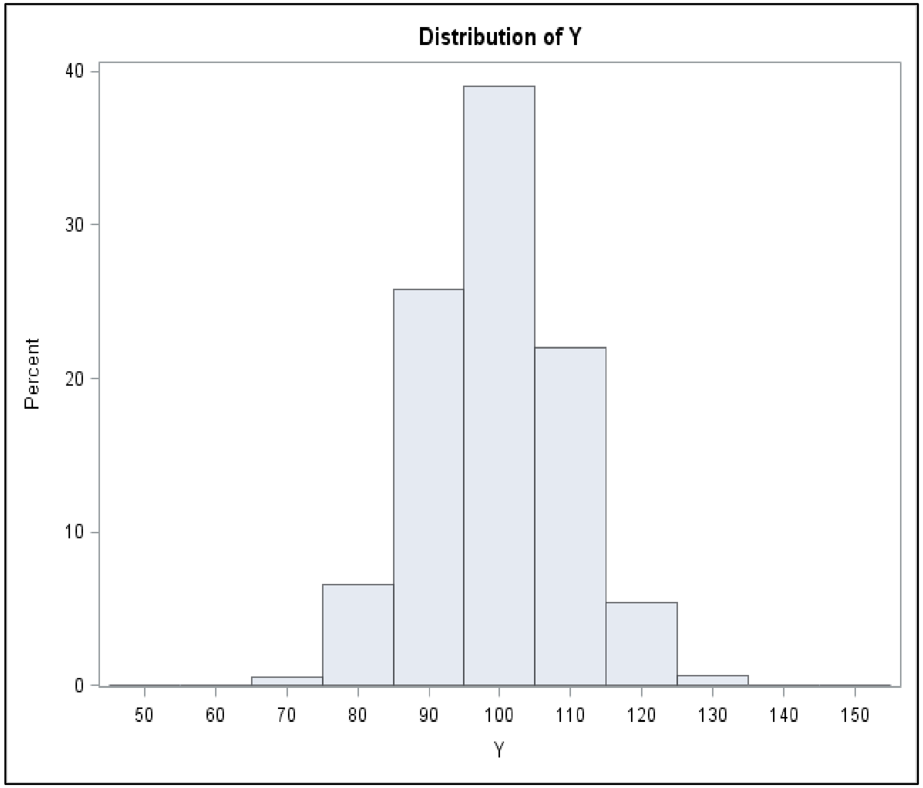
→ "Start at $i = 1$, subtract $\bar{y}$ from each y value, square that result, sum until N, then divide by $N - 1$"

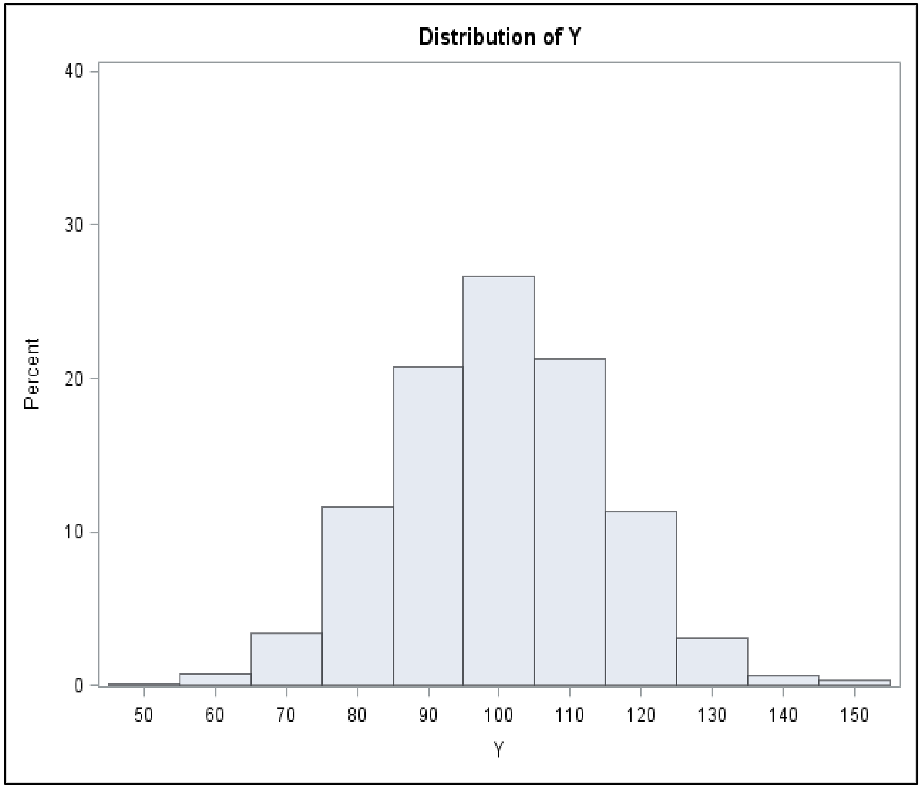
- Squaring is necessary to maintain absolute magnitudes, but because squared units are less interpretable than raw-data units, the standard deviation (SD, the square root of variance) can be more intuitive: **SD is the <u>average distance</u> for any given observation from the mean** (i.e., SD in results describes a variable's dispersion across persons)

- Btw, in the denominator of variance, $\boldsymbol{N-1}$ is used instead of N to adjust for needing the sample mean in order to calculate the sample variance; later on this term will be called "**denominator degrees of freedom (DF)**"

Illustrating Differences in Dispersion (Mean = 100 in both histograms)

Standard Deviation (SD) = **10**, Variance (VAR) = SD*SD = 100	Standard Deviation (SD) = **15**, Variance (VAR) = SD*SD = 225





Example Variable for Income: Get Mean, SD, and Variance

In **SAS**, using **PROC MEANS**:

```
PROC MEANS NDEC=3 N MEAN STDDEV VAR MIN MAX
  DATA=work.Example1;
VAR income;
RUN;
```

Analysis Variable : income Annual Income in 1000s					
N	Mean	Std Dev	Variance	Minimum	Maximum
734	17.303	13.792	190.209	0.245	68.600

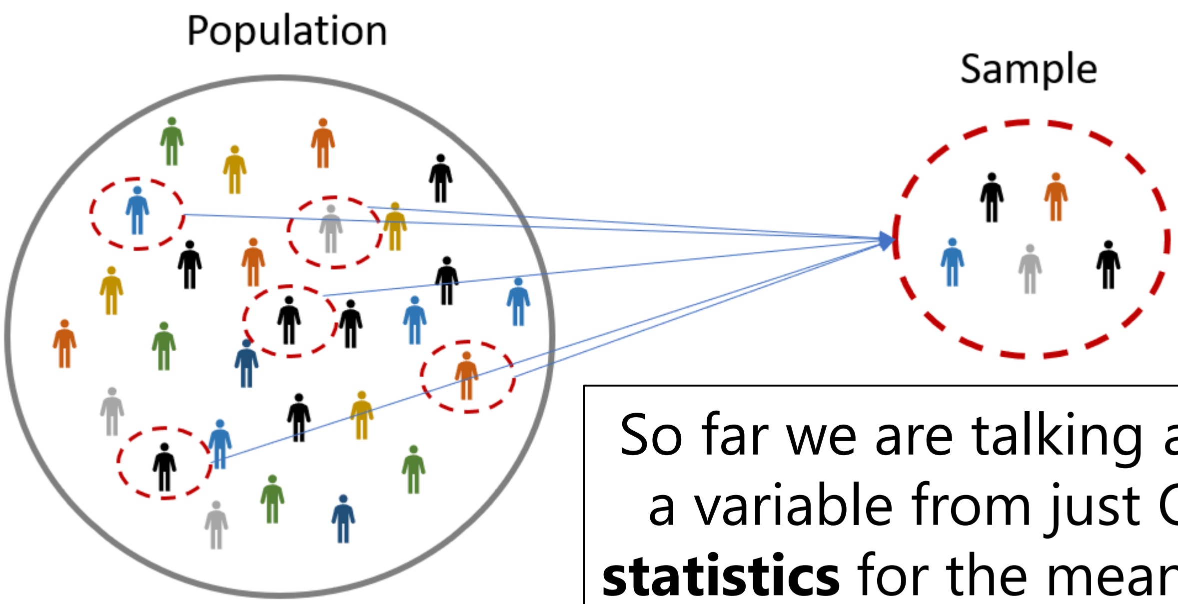
In **STATA**, using the command **SUMMARIZE** (add option DETAIL to get variance, too):

```
summarize income
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      income |       734    17.30287    13.79163       .245       68.6
```

From The Population To A Sample…

- To what **population** do we want to make inferences?
  - Numeric characteristics of the population are called "**parameters**"
- By what process should we select our **sample**?
  - Numeric characteristics of the sample are called "**statistics**"



So far we are talking about summarizing a variable from just ONE sample using **statistics** for the mean and variance—but those statistics are supposed to reflect the **parameters** of the intended population

Image borrowed from: https://www.sigmamagic.com/blogs/online-sample-size-calculators/

Sample vs. Population Notation for the Mean and Variance

- **Mean** (M) = average = central tendency = first "moment"
  - $\boldsymbol{\mu}$ ('mu") for the population is estimated by $\bar{\boldsymbol{y}}$ ("y bar") from a sample

- **Variance** (VAR) = squared dispersion = second "moment"
  - $\boldsymbol{\sigma^2}$ ("sigma squared") for the population is estimated by $\boldsymbol{s^2}$ from a sample
  - **Squared average deviation** of any given person from the mean
  - Squaring prevents ± deviations from mean from cancelling each other out

- **Standard deviation** (SD) = dispersion= square root of variance
  - $\boldsymbol{\sigma}$ ("sigma") for the population is approximated by $\boldsymbol{s}$ from a sample
  - **Average deviation** of any given person from the mean (in data units)

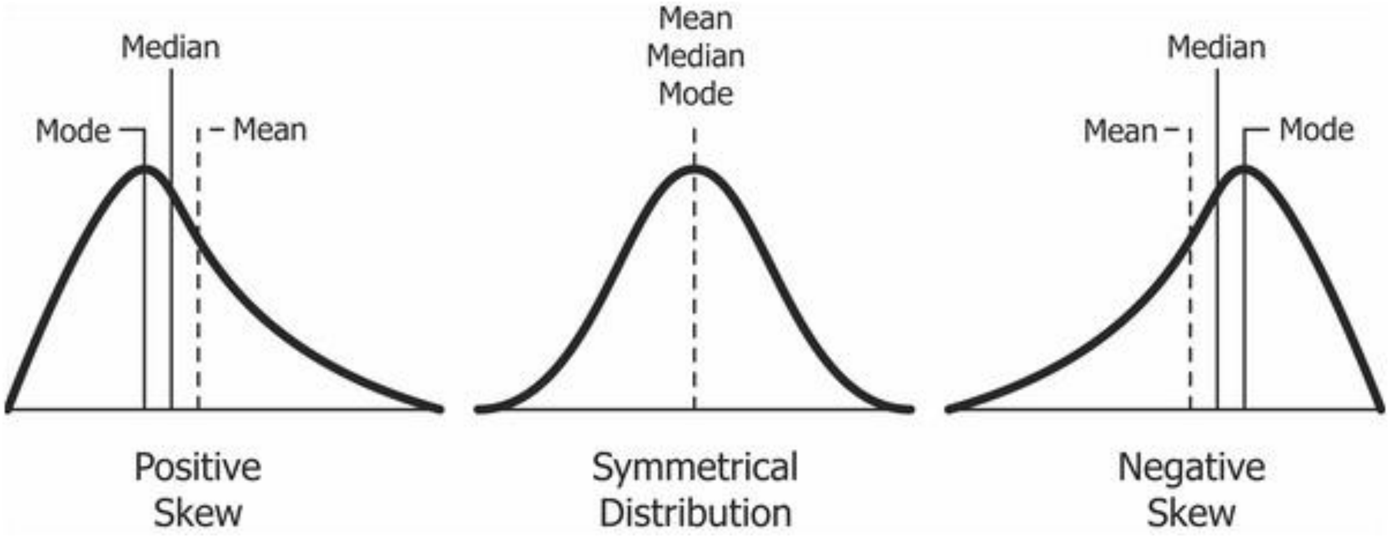
- Also sometimes reported: "coefficient of variation" = SD / mean

Salient Feature #3 of Quantitative Variables: Skewness (Asymmetry)

- **Skewness** (third "moment") follows a similar pattern:

$$\text{Skewness} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{s}\right)^3$$

→ Skewness will be 0 if the variable is symmetric(al)



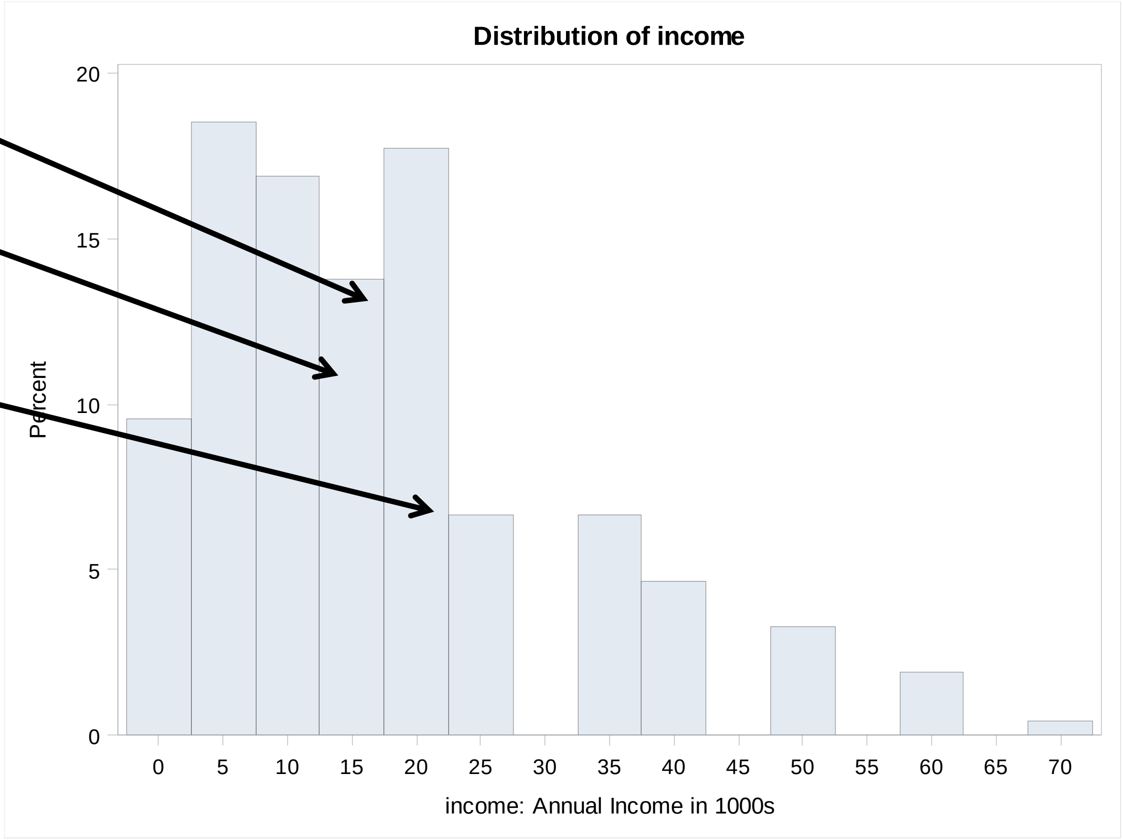
Note: Mean, median, and mode will diverge in asymmetric variables, so which one you report then matters!

Named direction of skew is where <u>the tail</u> is headed!

Image borrowed from: https://www.kullabs.com/classes/subjects/units/lessons/notes/note-detail/9958

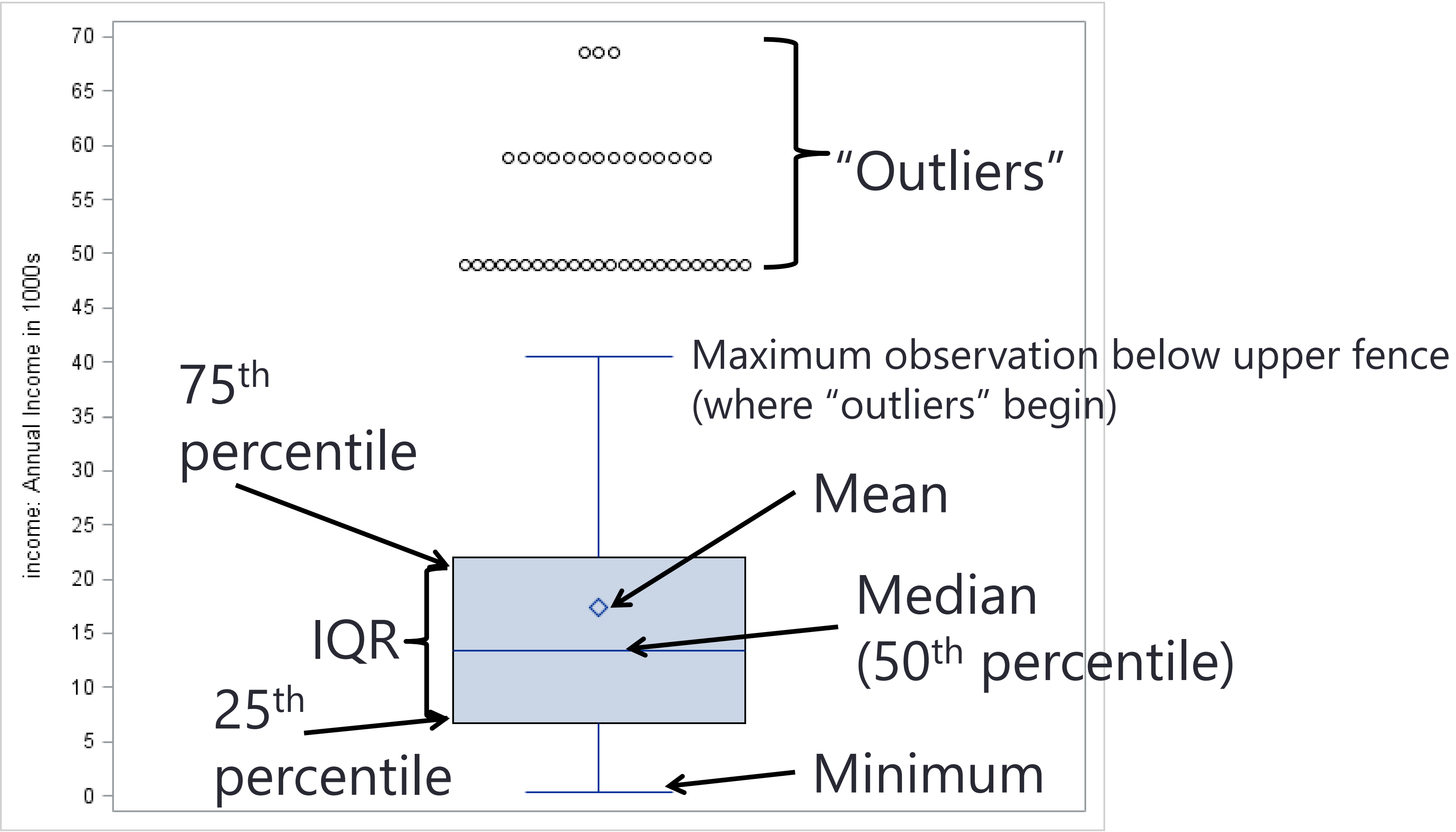
Example: Skewness in Income

- **Central tendency:**
  - Mean (M) = 17.31
  - Median = 13.48
    - Btw, = 50th percentile
  - Mode = 22.05
- **Dispersion:**
  - $VAR = SD^2$ = 190.21
  - SD = 13.79
  - Inter-quartile range:
    - IQR = 75th − 25th percentiles
    - IQR = 22.05 − 6.74 = 15.31



Should also report the **range**: the **minimum** and **maximum** values (0.245 and 68.60 here)

Summarize (Asymmetric) Quantitative Variables using a "**Box Plot**"



Get These Additional Statistics: In SAS using PROC UNIVARIATE

```
PROC UNIVARIATE DATA=work.Example1; VAR income; RUN;
```

Moments			
N	734	Sum Weights	734
Mean	17.3028747	Sum Observations	12700.31
Std Deviation	13.7916296	Variance	190.209048
Skewness	1.16073362	Kurtosis	1.10205445
Uncorrected SS	359175.104	Corrected SS	139423.232
Coeff Variation	79.7071579	Std Error Mean	0.50905834

Quantiles (Definition 5)	
Level	Quantile
100% Max	68.6000
99%	58.8000
95%	49.0000
90%	40.4250
75% Q3	22.0500
50% Median	13.4750
25% Q1	6.7375
10%	2.6950
5%	0.9800
1%	0.2450
0% Min	0.2450

Basic Statistical Measures			
Location		Variability	
Mean	17.30287	Std Deviation	13.79163
Median	13.47500	Variance	190.20905
Mode	22.05000	Range	68.35500
		Interquartile Range	15.31250

Get These Additional Statistics: In STATA using SUMMARIZE (DETAIL)

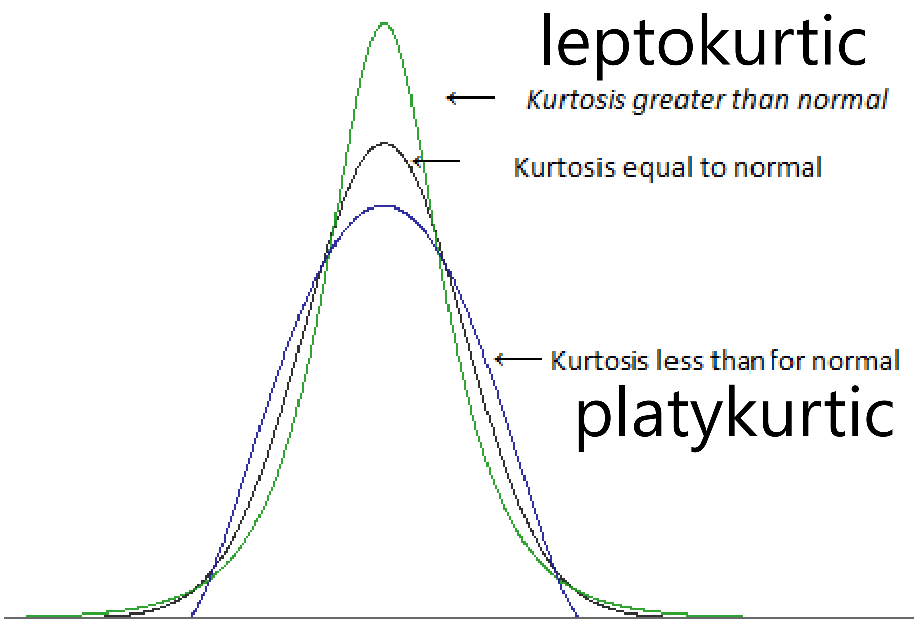
```
summarize income, detail

              income: Personal Income in 1000s
-------------------------------------------------------------
      Percentiles      Smallest
 1%         .245           .245
 5%          .98           .245
10%        2.695           .245       Obs                 734
25%       6.7375           .245       Sum of Wgt.         734

50%       13.475                      Mean           17.30287
                        Largest       Std. Dev.      13.79163
75%        22.05           58.8
90%       40.425           68.6       Variance        190.209
95%           49           68.6       Skewness        1.15836
99%         58.8           68.6       Kurtosis       4.086398
```

Btw, One More Feature of Quantitative Variables: Kurtosis

- **Kurtosis** (fourth "moment") follows a similar pattern:

$$\text{Kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{s}\right)^4 - 3$$

→ Will be 0 if the variable is symmetric(al)

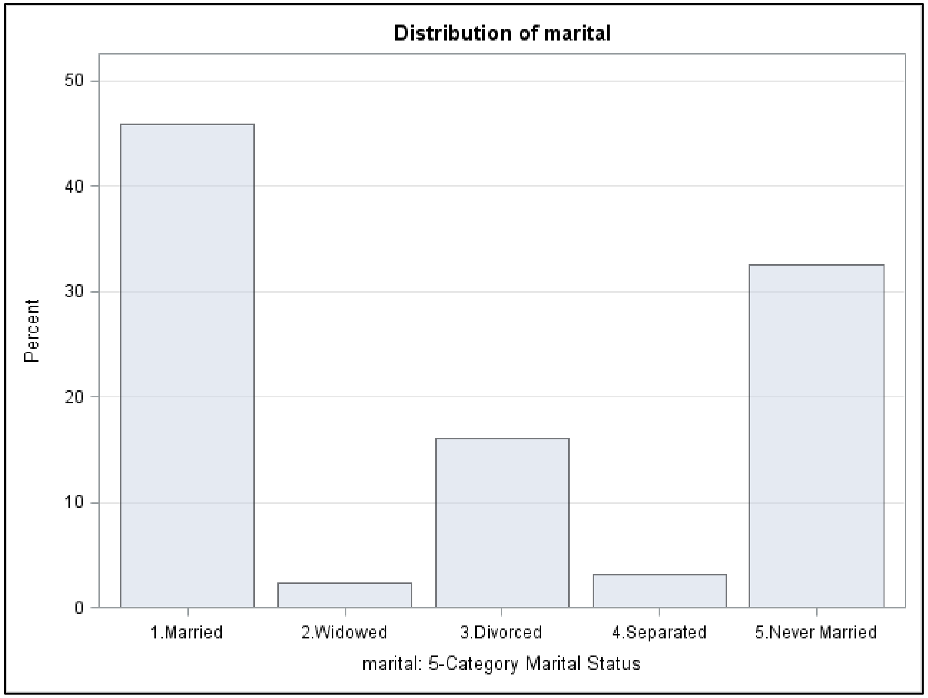


Note: Extent of kurtosis is hard to differentiate from variance in real data, so don't worry about this one

Image borrowed from: https://stats.stackexchange.com/a/143522/124771

Means for Categorical Variables?

- For binary variables coded 0 or 1, **the mean** is calculated the same way but it **is called the "proportion"** instead
- For **nominal variables** with >2 options, **a single mean does not make sense!**
  - e.g., for nominal marital status, *M* = 2.74… ?!?
  - You may see means calculated for **ordinal variables** but they should give you pause….
    - e.g., 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree…. could also be 1, 20, 300, 4000, 50000



Variances for Categorical Variables?

- For **binary variables coded 0 or 1**, variance and skewness are not separate properties (as they are in quantitative variables)
  - If $\boldsymbol{p}$ = proportion of 1 values, and $\boldsymbol{q}$ = proportion of 0 values:
  - Mean $\overline{\boldsymbol{y}} = \boldsymbol{p}$, variance $\boldsymbol{s}^2 = \boldsymbol{p} * \boldsymbol{q}$, and skewness = $\frac{\mathbf{1}-\mathbf{2}\boldsymbol{p}}{\sqrt{\boldsymbol{p}*\boldsymbol{q}}}$

Mean and Variance of a Binary Variable

Mean ($\boldsymbol{p}$)	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

- For variables with >2 categories, **each pair of categories** would have its **own $\boldsymbol{p}$ and $\boldsymbol{q}$** (and thus variance/skewness)
  - So the **percent for each category is enough to report** (i.e., the pairwise variance and skewness values are not helpful)

Intermediate Summary

- What kind of **univariate summary statistics** are relevant to report depends on the type of variable to be described:

  - **Quantitative variables (numbers are numbers):**
    - If "symmetric enough": Min, Max, Mean, SD (or SD^2 = variance)
    - If not, add median (for central tendency) and IQR (for dispersion) that are "robust" to outliers (extreme values) or general skewness
    - Binned-value histograms or boxplots (or violin plots) make good visuals

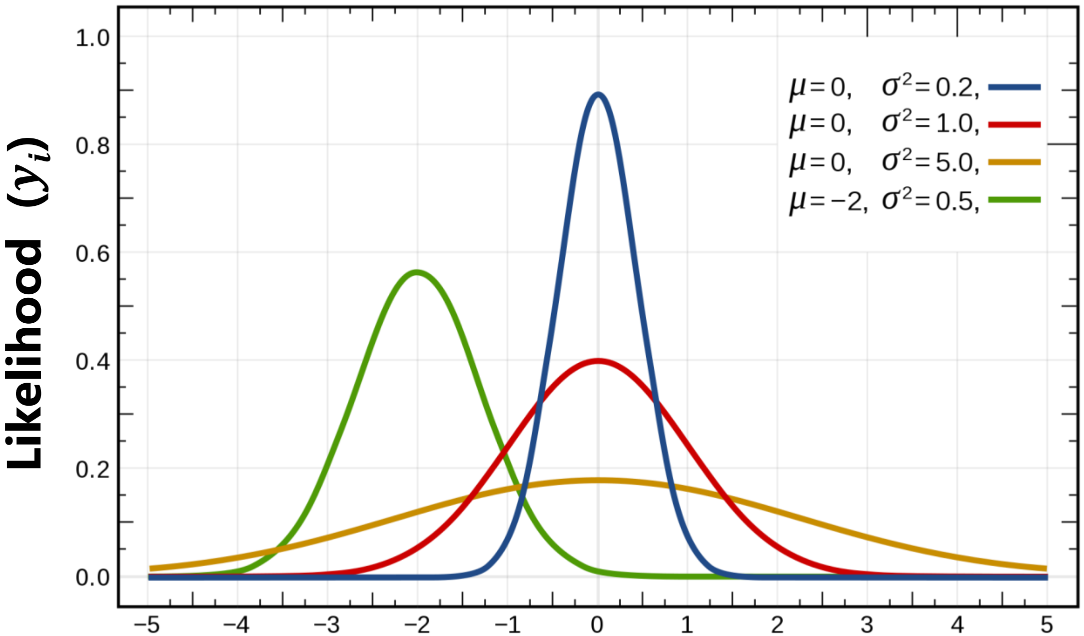
  - **Categorical variables (numbers are just labels):**
    - **Binary** (0 or 1): Mean (= **proportion** of 1 values); variance and skewness are then determined by the mean (i.e., they are redundant)
    - **Ordinal** or **Nominal** with **3+ categories**: **percentage** of each category; a single mean (or variance or skewness) makes no kind of sense
    - You may see ordinal variables treated as quantitative, but keep in mind this assumes real distances between the numbers used as labels
    - Bar graphs of the percentage in each category make a good visual

From Descriptive to Inferential Statistics

- So far we have considered examples of descriptive statistics, whose job is to summarize variables from a single sample
  - "**Descriptive**" → used to describe, condense, or summarize one sample
- But if we want to generalize from our one sample back to the intended population, we then also need inferential statistics
  - "**Inferential**" → used to make statements about population values

- Inferential statistics rely on **Probability Distribution Functions (PDFs)**: mathematical equations that provide the *likelihood* of the possible values of a variable of that type (abbreviated "distributions")
  - For **discrete** variables (integers only), PDFs provide the **probability of any exact value** (="probability mass function" or PMF)
  - For **truly continuous** variables, the probability of any one value is undefined, so PDFs provides the probability over a range of values instead; "**probability**" **switches to** "**likelihood**" (but is the same idea)

A common PDF: Normal Distribution (or "Gaussian" or "bell curve")



Univariate Normal PDF:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left[-\frac{1}{2} * \frac{(y_i-\mu)^2}{\sigma^2}\right]$$

Two parameters:

- $\boldsymbol{\mu}$ = "mu" = mean
- $\boldsymbol{\sigma^2}$ = "sigma squared" = variance
- Ranges from ±∞ (so is actually continuous)

Is **symmetric**, so:

- skewness = 0
- One middle: mean = median = mode

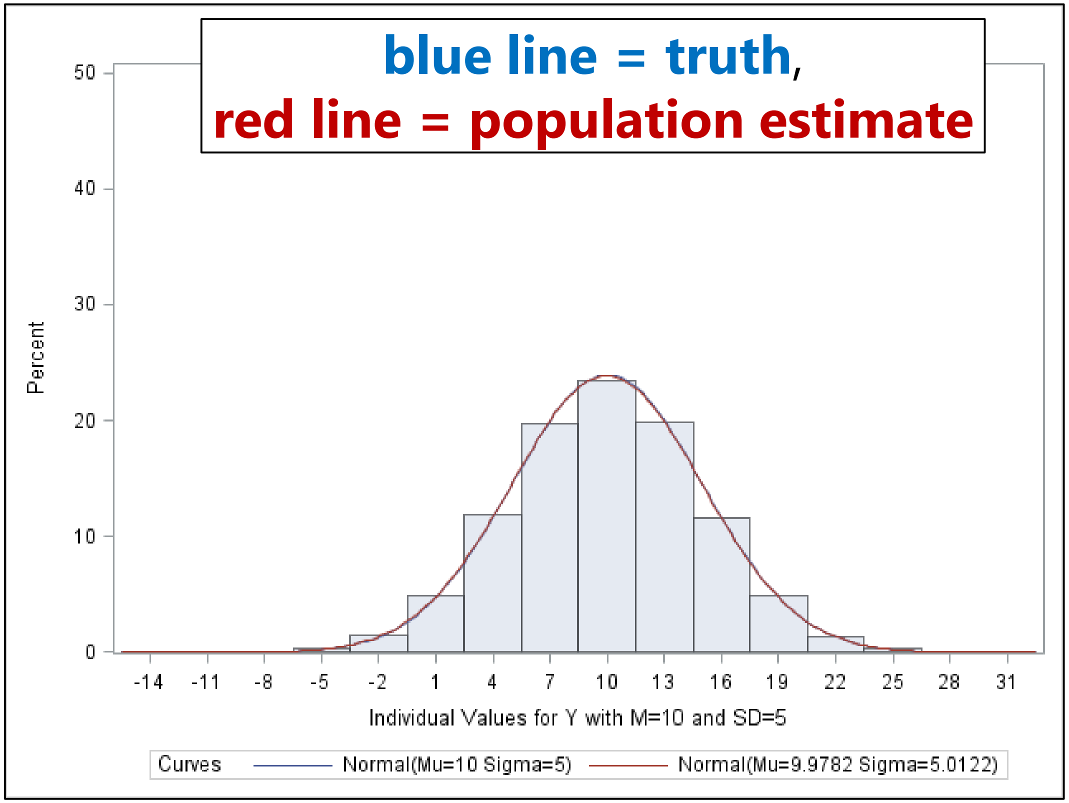
Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

From Descriptive to Inferential Statistics

- For a "**univariate**" analysis (i.e., about one variable) to be able to make **inferences to the population** from a single sample:
  - At a minimum, this involves **indexing the inconsistency** of the sample-specific summary statistic, e.g., of the sample mean $\bar{y}$
    - i.e., if you repeated the same study, **how close** would the mean of the new sample be to the mean of the current sample?
    - Index of inconsistency can be used to form an **expected range** in which the statistic would be found across repeated samples
  - Could also involve a **comparison** of the sample-specific summary **statistic** to an expected **population value**
    - e.g., how different is sample mean $\bar{y}$ from the population mean μ?
    - Said differently, if the population mean really were true, **how likely** are we to have observed the sample mean that we found?

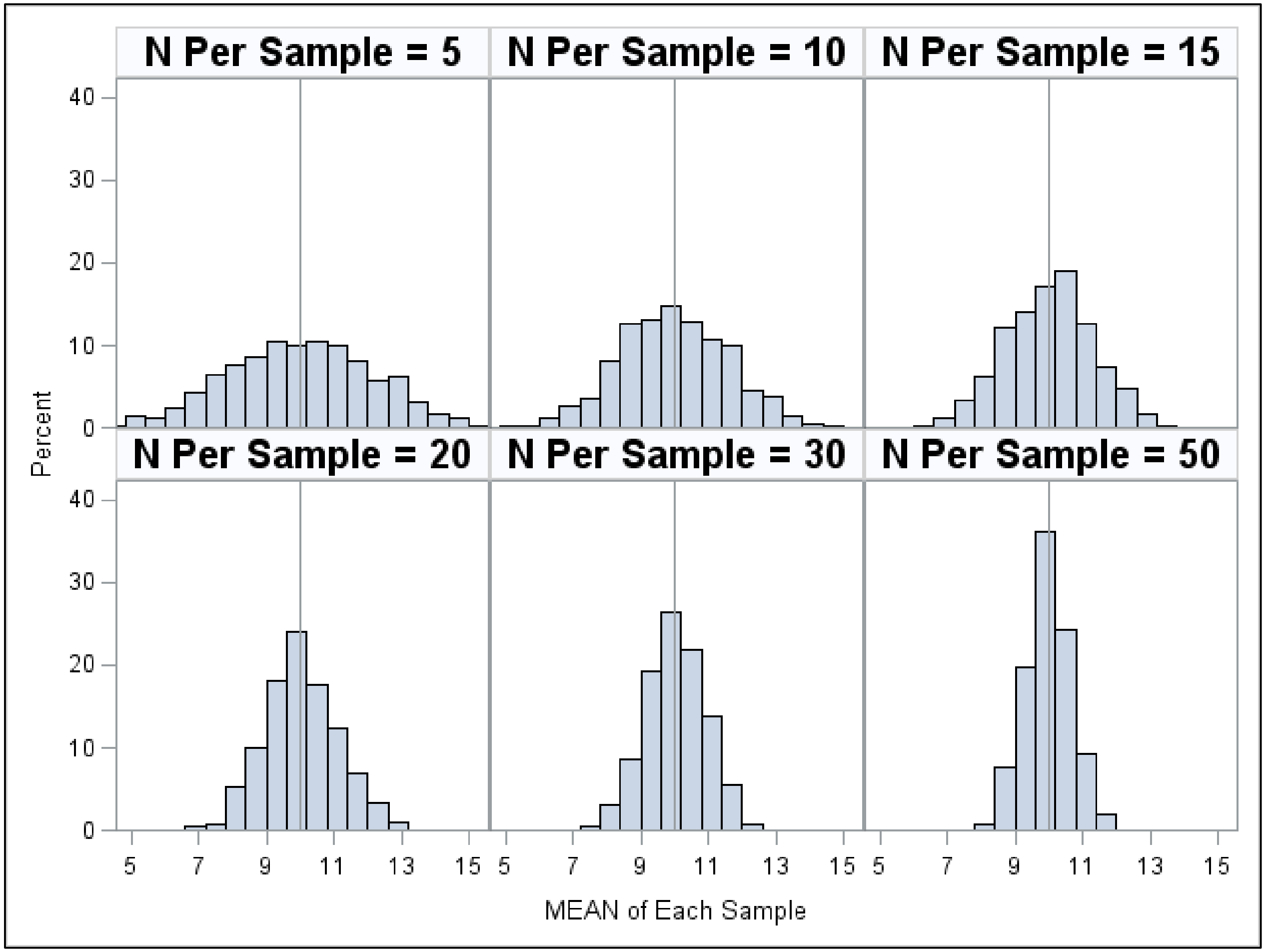
Building Intuition about Sampling Distributions of Statistics (the mean for now)

- **What affects how close $\bar{y}$ is to the true value of μ?**
- Demo: I made my own quantitative variable* y_i in a population of 100,00 fake people
  - Population mean: $\mu = 10$
  - Population VAR: $\sigma^2 = 25$
  - So y_i is off the mean by $SD = 5$ on average



* Used a "**normal**" distribution here to generate y_i (as described earlier)

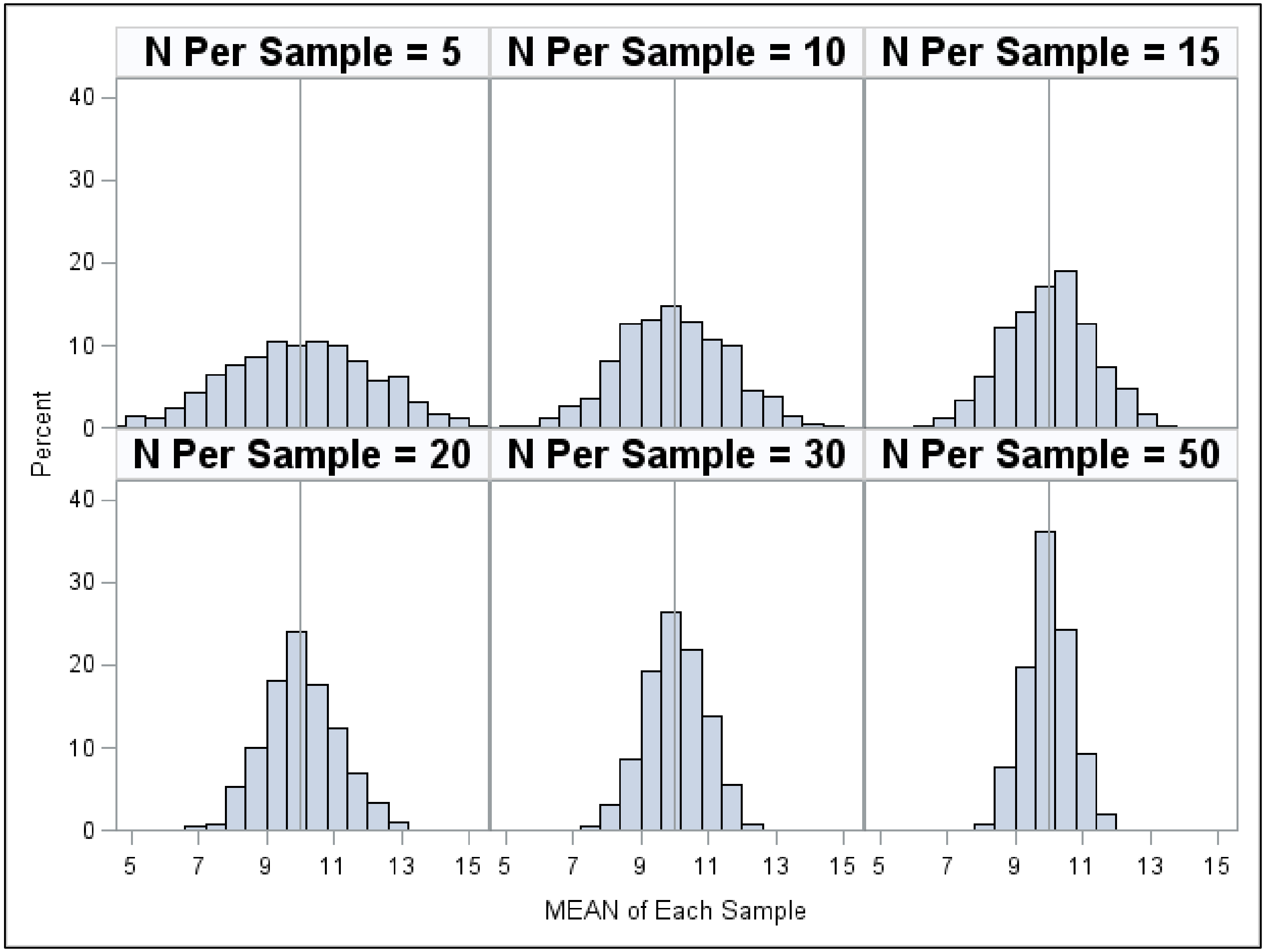
1000 samples each for different N…



Note: These bars do not show individual people!
They are summaries for **distinct samples** of people.

- Population values: Mean $\boldsymbol{\mu} = \mathbf{10}$ (SD $\boldsymbol{\sigma} = \mathbf{5}$)
- Histograms show **differences across samples** in each sample's **mean** ($\bar{y}_s$)
- These depict the N-specific "**sampling distribution**" of $\bar{y}_s$
- **More N** in each sample → **less dispersion in $\bar{y}_s$** across samples (**more consistency**)

1000 samples each for different N…



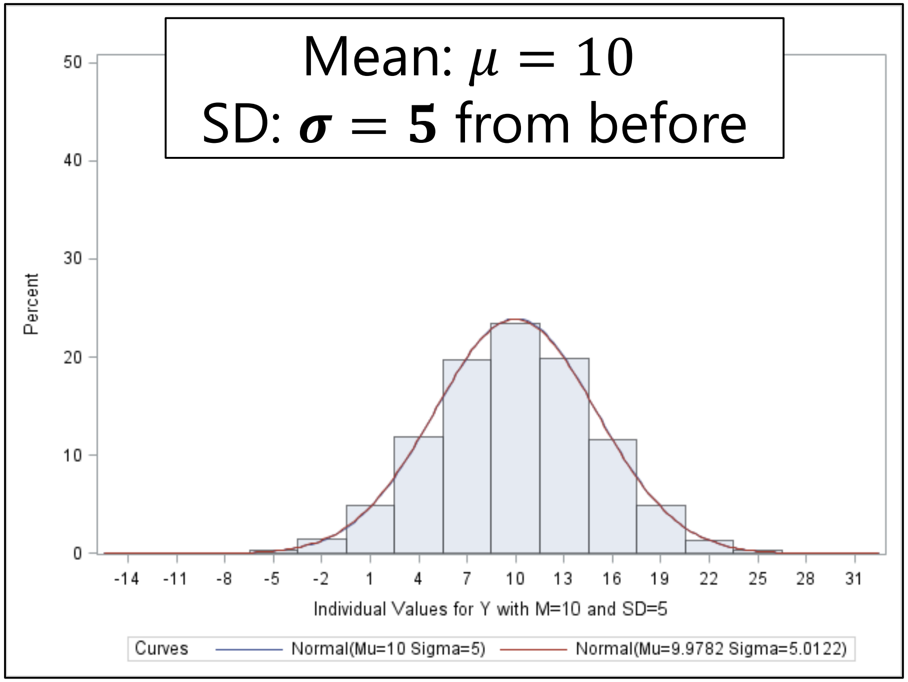
Note: These bars do not show individual people!
They are summaries for **distinct samples** of people.

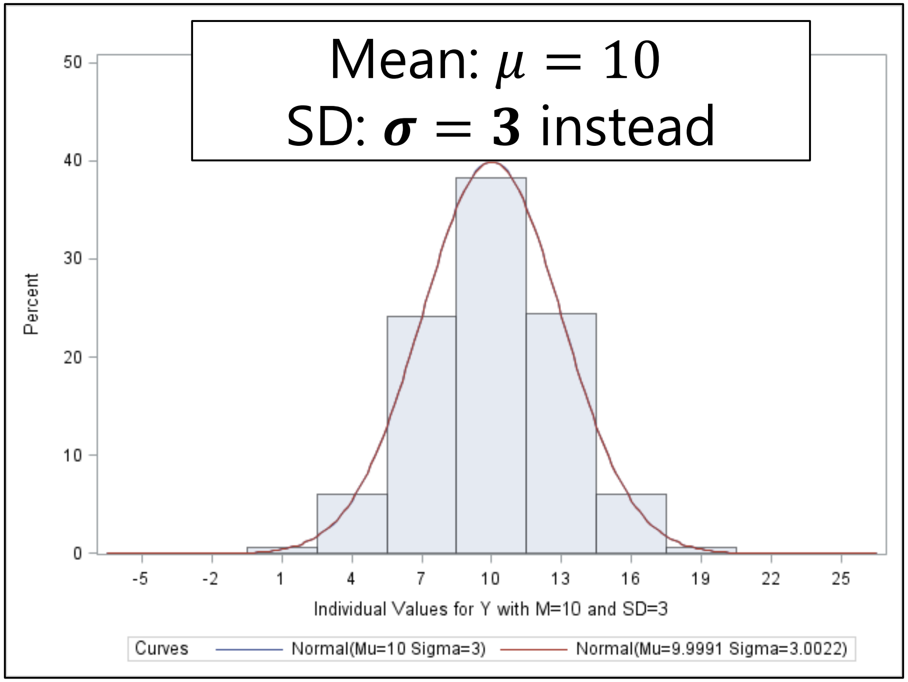
- Population values:
  Mean $\boldsymbol{\mu} = \mathbf{10}$
  (SD $\boldsymbol{\sigma} = \mathbf{5}$)
- **More N → less SD in $\bar{y}_s$** across samples

N Per Sample	Mean $\bar{y}_s$	SD $\bar{y}_s$
5	9.97	2.17
10	9.98	1.60
15	10.00	1.28
20	10.03	1.08
30	10.03	0.89
50	9.97	0.69

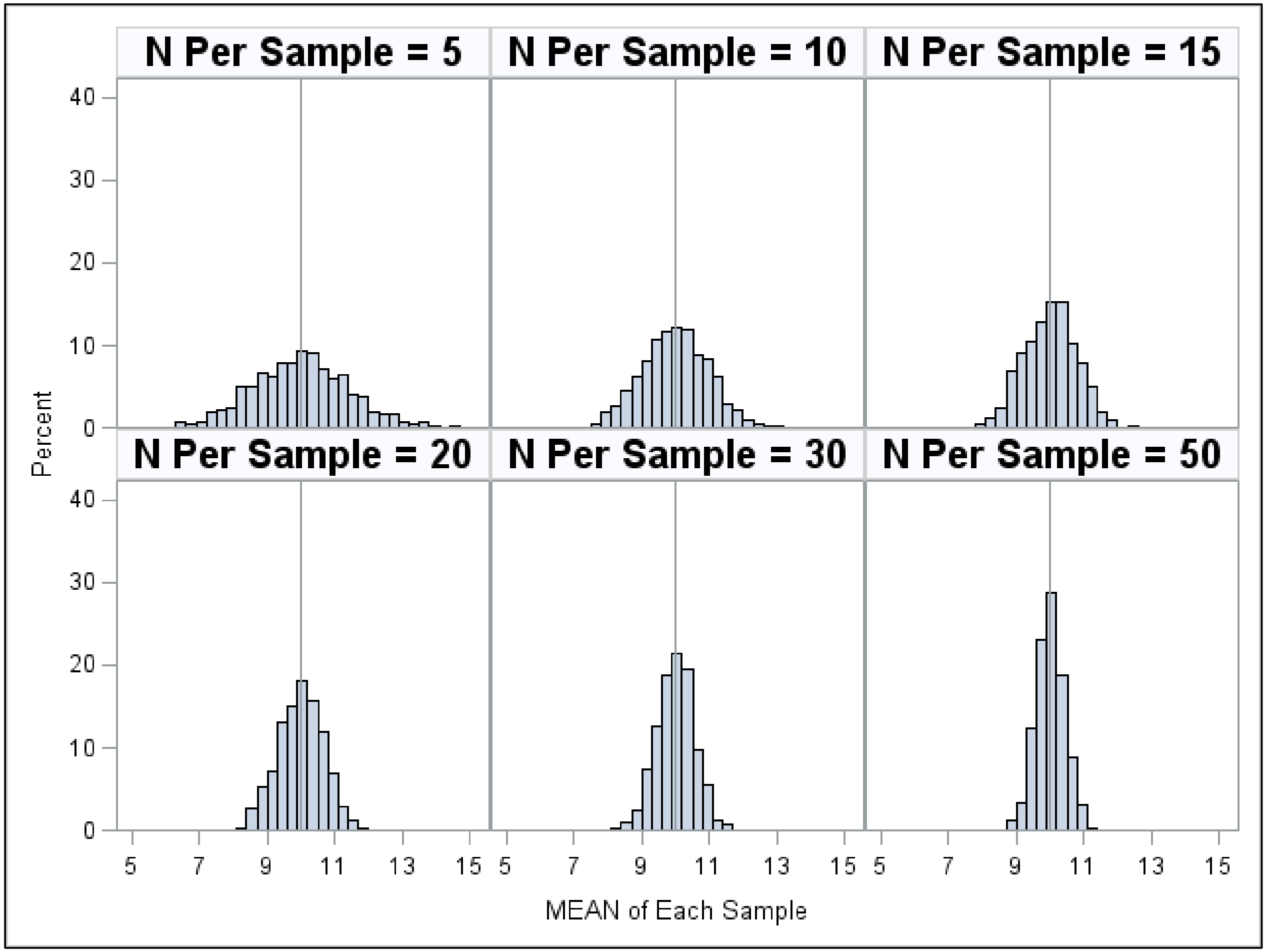
Building Intuition about Sampling Distributions of Statistics (the mean for now)

- So **sample size N** improves the consistency of the mean $\bar{y}_s$ for any sample
  - As within-sample N **increases**, sample mean $\bar{y}$ **will be closer to μ on average**
- What else affects precision of $\bar{y}_s$? How **persons vary from each other**!





1000 samples each for different N…

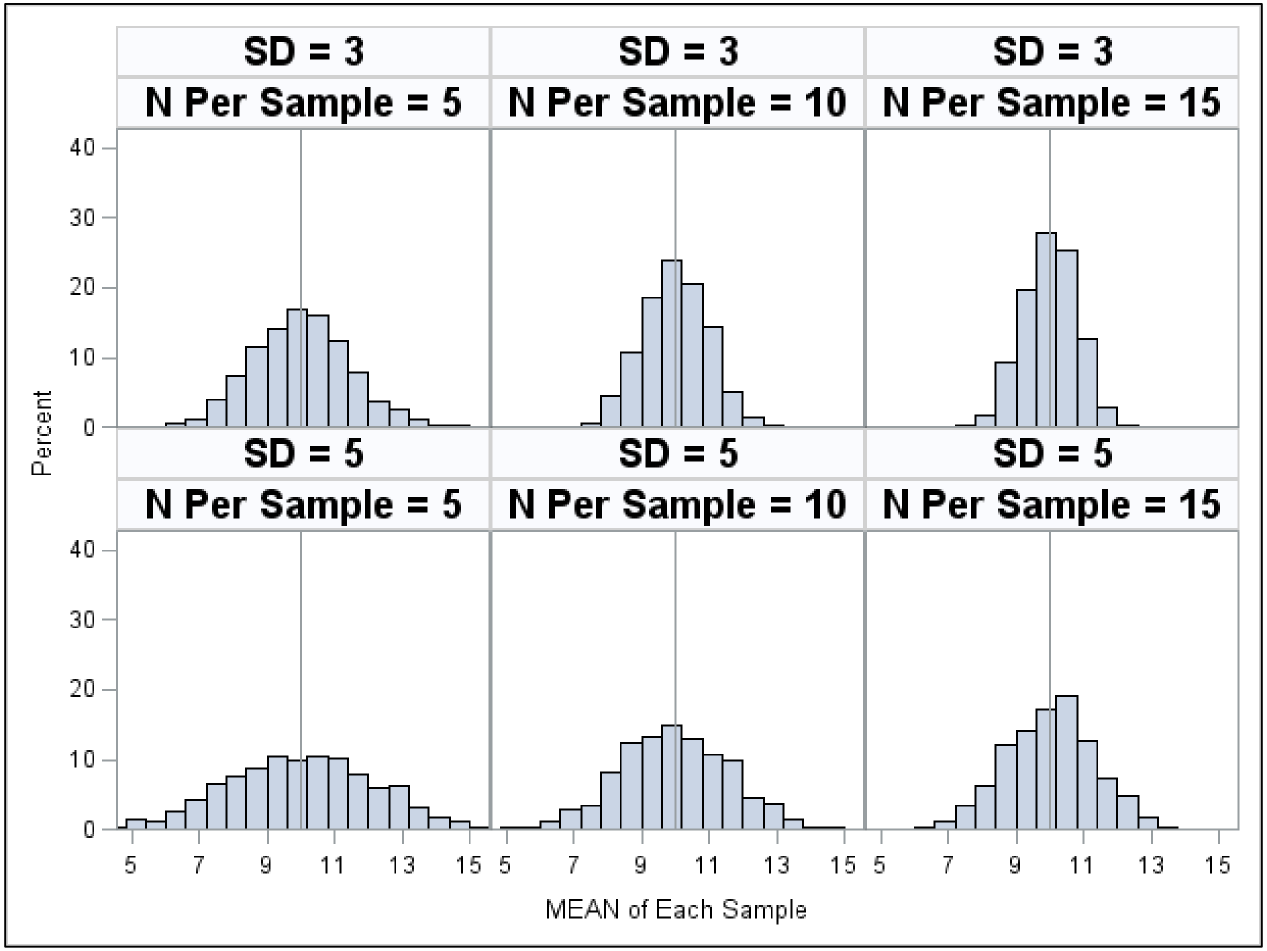


These bars still do not show individual people!
They are summaries for **distinct samples** of people.

- Population values:
  Mean $\boldsymbol{\mu} = \mathbf{10}$
  (SD $\boldsymbol{\sigma} = \mathbf{3}$ now)
- **More N → less SD in $\bar{y}_s$** across samples

N Per Sample	Mean $\bar{y}_s$	SD $\bar{y}_s$
5	10.01	1.42
10	10.00	0.96
15	10.01	0.78
20	9.99	0.67
30	10.00	0.56
50	10.00	0.42

Effects of N and SD on Precision of $\bar{y}_s$



These bars still do not show individual people!
They are summaries for **distinct samples** of people.

Left to right:

- **More N** in each sample → **less dispersion in $\bar{y}_s$** across samples

Top to bottom:

- **More SD** in each sample → **more dispersion in $\bar{y}_s$** across samples

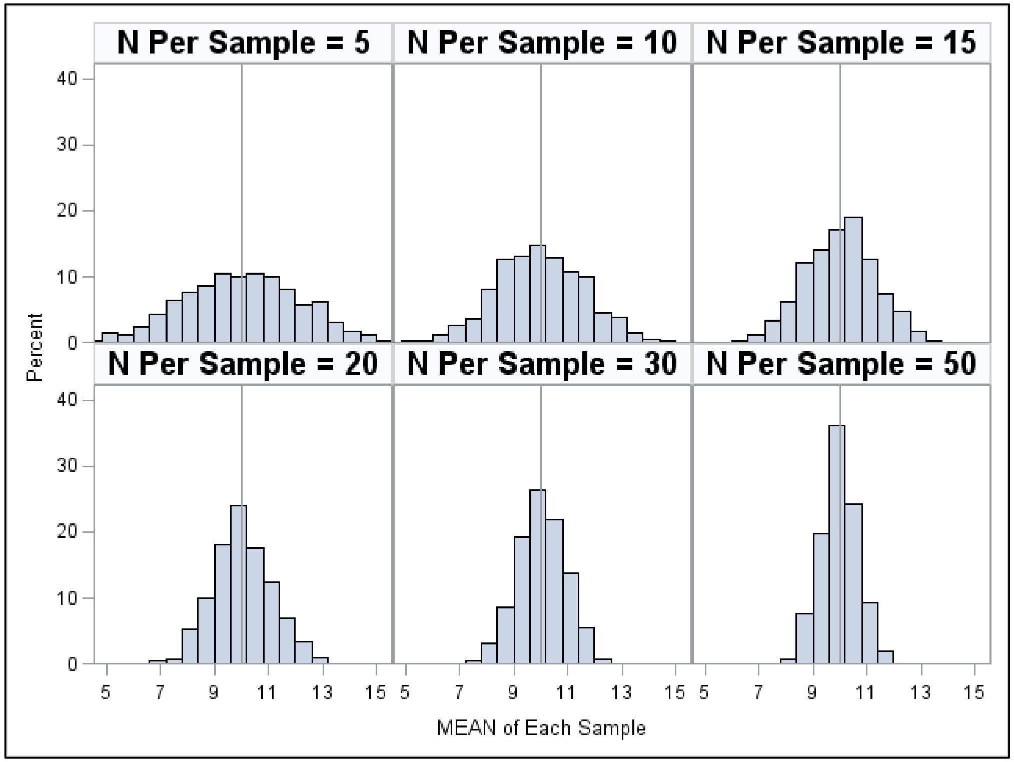
Anticipating Precision of Sample Mean $\bar{y}_s$

- In the example from the previous slides, we had a **known finite population** from which multiple random samples were selected
  - **Inconsistency of $\bar{y}_s$** could be indexed by standard deviation (SD) across samples → **more N, less variance → smaller SD of $\bar{y}_s$** (more consistent)

- Given only one sample, we can still **anticipate the SD of $\bar{y}$**:
  - **SD of $\bar{y}_s$ across samples ← Standard Error of the Mean** = $SE = \frac{\sigma}{\sqrt{N}}$
    - Note that SE includes the population SD σ, which must be replaced by the sample-estimated SD s when σ is unknown (i.e., most of the time)
  - **SE of the mean** is the expected average deviation of any given *sample mean* $\bar{y}$ from the *population mean* μ (even if you do not know μ)
    - Is NOT the same as SD of y_i (s) which is the average deviation of any given *observation* (i.e., person) from the *sample mean* (that you can calculate)
    - In general, the term **"SE" refers to the SD of a statistic's sampling distribution** (e.g., how the variance differs across samples is also described by its SE)

SE of Mean Predicts SD of $\bar{y}_s$

Population values for y_i variable: Mean $\boldsymbol{\mu} = \mathbf{10}$, SD $\boldsymbol{\sigma} = \mathbf{5}$



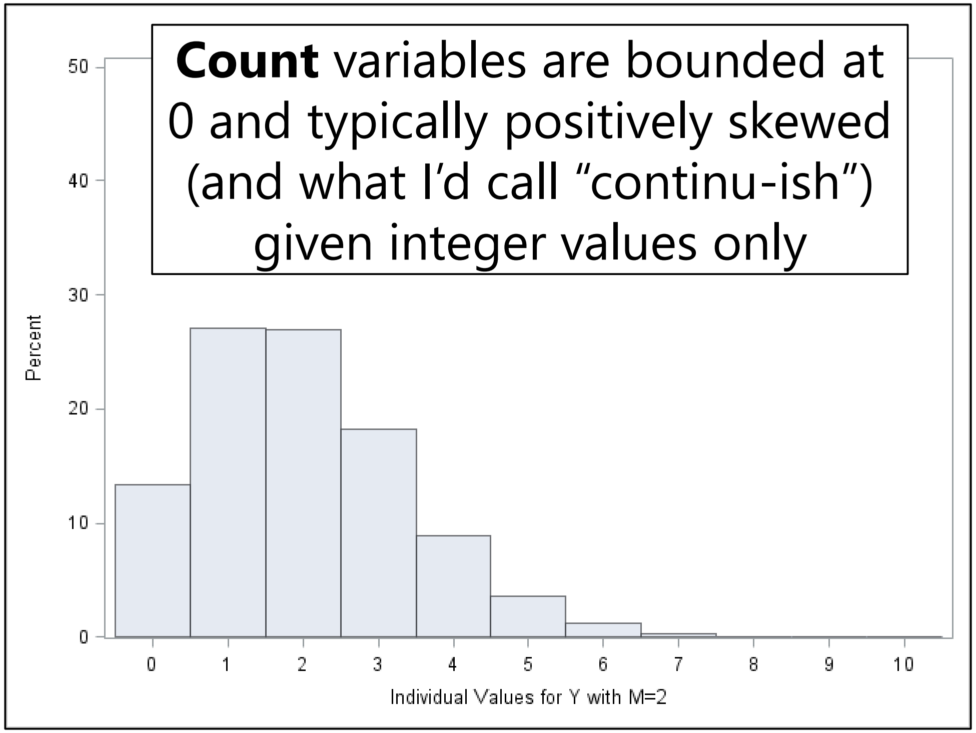
N	Mean $\bar{y}_s$	SD $\bar{y}_s$	Mean SE with: σ	Mean SE with: s
5	9.97	**2.17**	2.24	2.13
10	9.98	**1.60**	1.58	1.55
15	10.00	**1.28**	1.29	1.28
20	10.03	**1.08**	1.12	1.11
30	10.03	**0.89**	0.91	0.91
50	9.97	**0.69**	0.71	0.71

The greater the sample size N, the better the estimate of each sample's SD, and the less it matters that SE is formed with sample SD (s) instead of the population SD (σ). But this distinction will matter more in smaller samples….

What about Other Kinds of Variables?

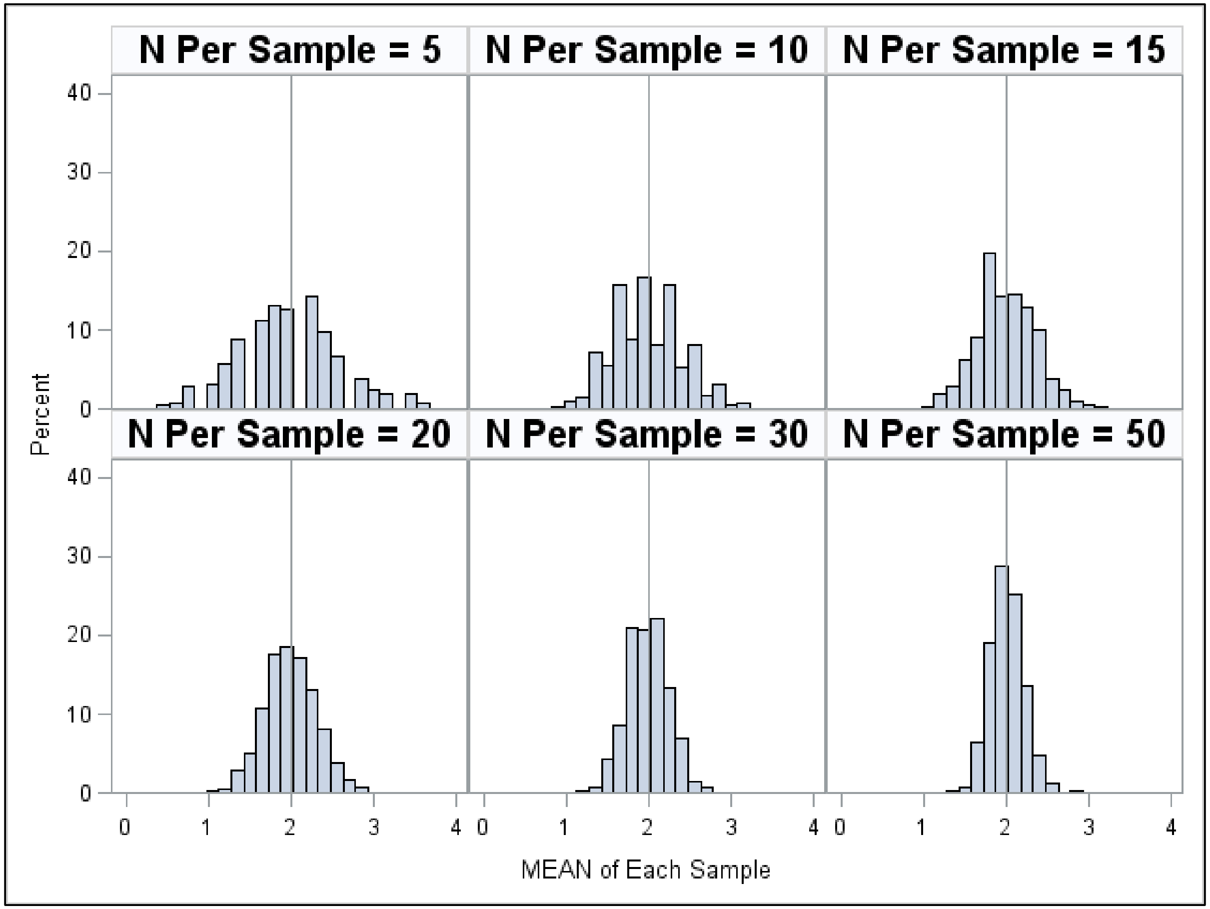
- It turns out **with more N** the **sampling distribution of $\bar{y}_s$ becomes more normal** *no matter what the observed variable's distribution is*
  - Btw: More N → more normal $\bar{y}$ distribution → is "Central Limit Theorem"

- Demo: I simulated a **count variable*** $\boldsymbol{y_i}$ in a population of 100,00 fake people
  - Population mean: $\boldsymbol{\mu = 2}$
  - Population VAR: $\boldsymbol{\sigma^2 = 2}$
  - So $\boldsymbol{y_i}$ is off the mean by $\boldsymbol{SD} = \sqrt{\boldsymbol{2}}$ on average



* Used a "Poisson" distribution here to generate $\boldsymbol{y_i}$ (in which $\boldsymbol{\mu = \sigma^2}$)

1000 samples each for different N…



Note: The observed SD for the sampling distribution for $\bar{y}_s$: (a) is well-approximated by the mean SE for $\bar{y}_s$, and (b) appears normal, even for a count variable

- Population values:
  Mean $\boldsymbol{\mu} = \mathbf{2}$
  (VAR $\boldsymbol{\sigma}^2 = \mathbf{2}$)
- **More N → less SD in $\bar{y}_s$** across samples; $\bar{y}_s$ is also more normal

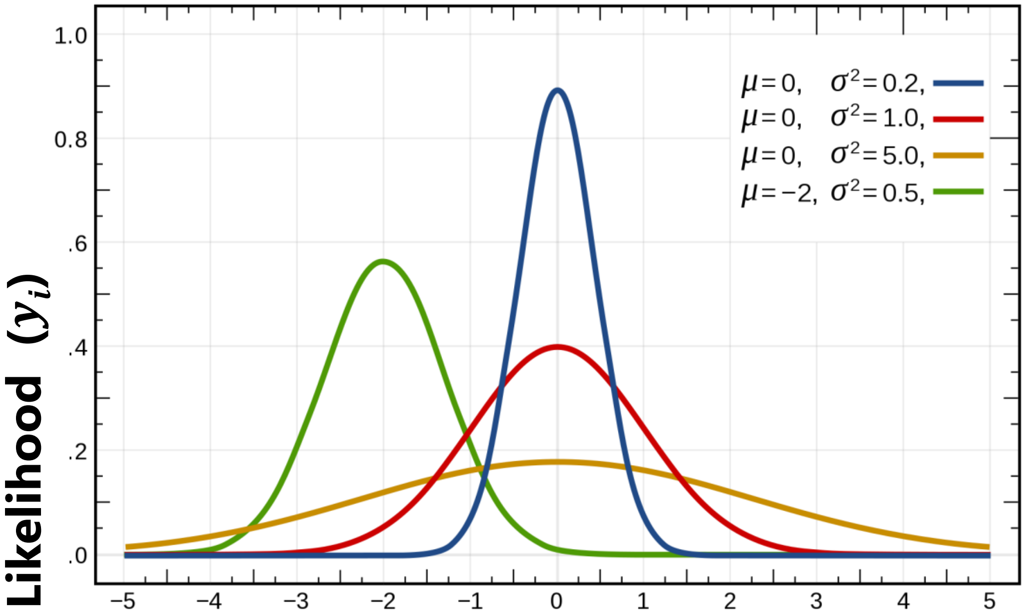
N	Mean $\bar{y}_s$	SD $\bar{y}_s$	Mean SE
5	1.98	0.61	0.59
10	1.99	0.42	0.43
15	1.99	0.35	0.36
20	1.99	0.31	0.31
30	1.99	0.25	0.25
50	2.01	0.20	0.20

Using the SE of the Mean to Make Inferences Back to the Population

- **SE of the mean** = average difference between a given sample mean $\bar{y}_s$ and the population mean μ (i.e., SE of the mean approximates the SD for the mean's distribution across repeated samples)
  - In general, **any sample statistic has an SE** for the statistic's average difference between a given sample value and its population value

- An SE can be used to express the **range of uncertainty** around a sample statistic (i.e., the mean here) across repeated samples by forming a **confidence interval**, which requires **two decisions**:
  - What **probability distribution function** can be used to describe the expected behavior of the statistic's sampling distribution?
    - Sample mean should become **normally distributed**, so let's start with that
  - **Level of confidence**: how often are you willing to be wrong?
    - Typical **confidence** level chosen is **95%**, so you'd be **wrong 5%** of the time
    - Btw, **wrong %** will be known as "**alpha level**" in hypothesis tests (stay tuned)

Standardizing the Normal Distribution…



Univariate Normal PDF:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} * \exp\left[-\frac{1}{2} * \frac{(y_i - \mu)^2}{\sigma^2}\right]$$

- Normal distribution uses an estimated mean and variance to **provide the likelihood of any y_i value**
- To make it useful for sample statistics (like the mean) for **variables on different scales**, we need a standardized version
- **The "z" metric with** $M = 0$ and $VAR = 1$ ($SD = 1$) creates a new "**standard normal distribution**"…

Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

Area Under Standard Normal Curve

y-axis created by:

$$f(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right)$$

The x-axis (called z_i) is in **standard deviation units** (where $SD = 1$)

Confidence Intervals using standard normal distribution have these z **"critical" values:**

90% within $z = \pm 1.65$

95% within $z = \pm 1.96$

99% within $z = \pm 2.58$

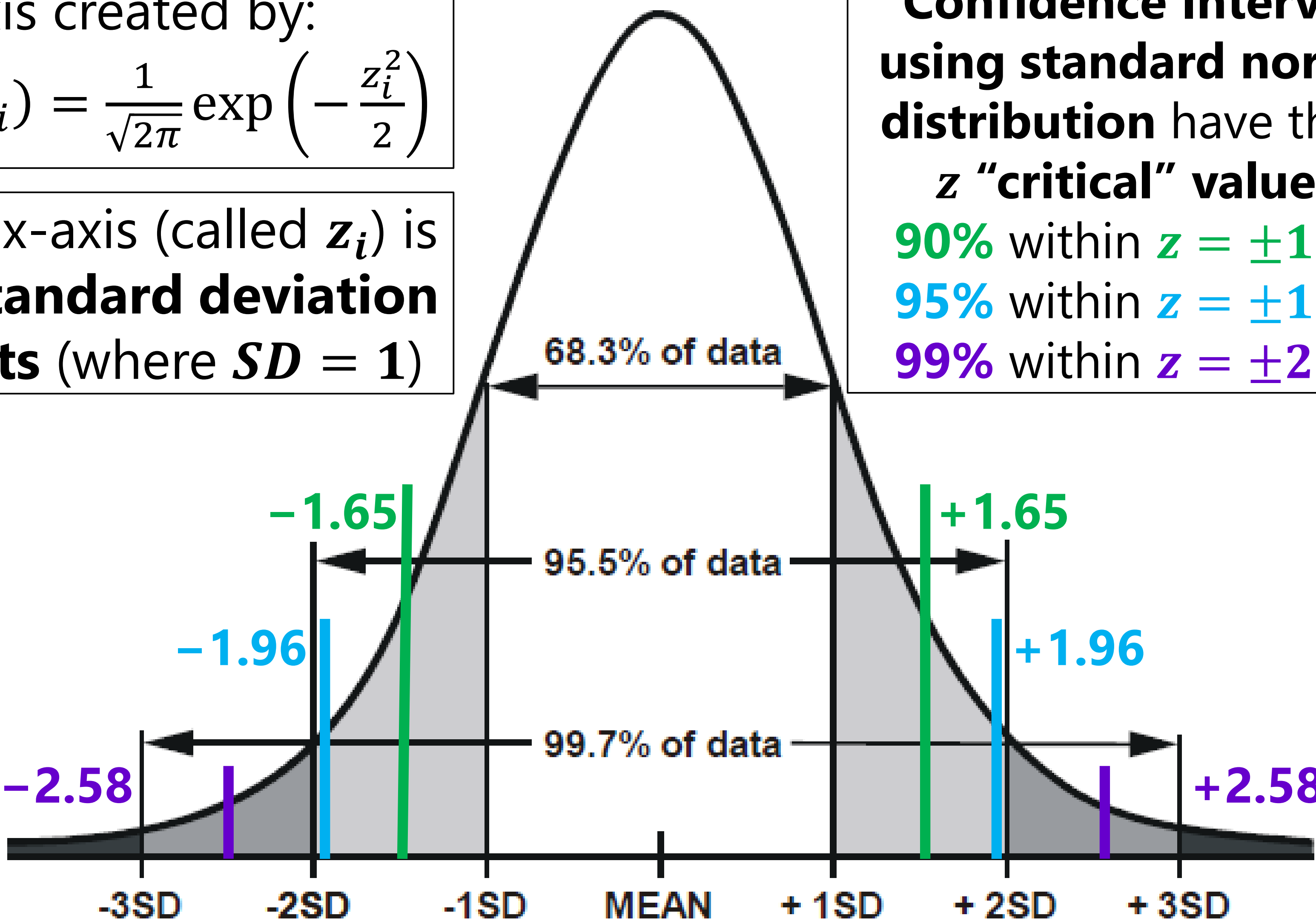
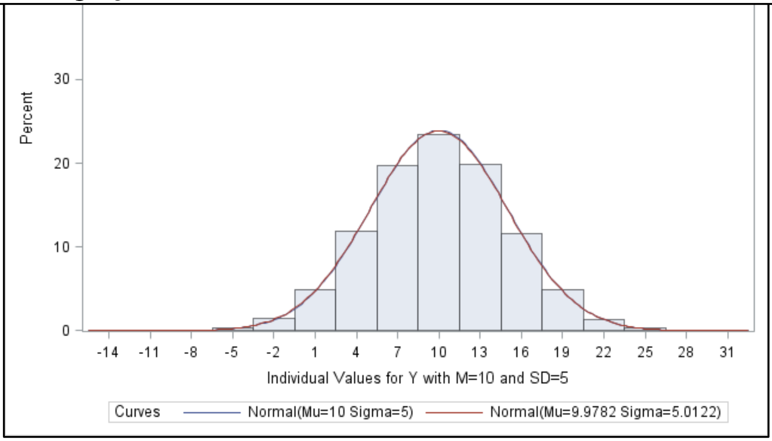


Image adapted from: http://my.ilstu.edu/~gjin/hsc204-hed/Module-5-Summary-Measure-2/Module-5-Summary-Measure-28.html

Confidence Interval for Sample Mean using z Standard Normal Distribution

- **That sample's statistics:**
  - Mean: $\bar{y} = 10.98$ (estimate)
  - SD: $s = 4.60$ (person dispersion)
  - SE of Mean $= \frac{s}{\sqrt{N}} = \frac{4.60}{\sqrt{50}} = 0.65$

Draw **1 sample** of $\boldsymbol{N} = \mathbf{50}$ from the $\boldsymbol{y_i}$ below with $\boldsymbol{\mu} = \mathbf{10}, \boldsymbol{\sigma} = \mathbf{5}$



- **Confidence Interval (CI):**
  $CI = Estimate \pm (critical * SE)$
  - **90%** CI for Mean: $CI = 10.98 \pm (\mathbf{1.65} * 0.65) = 9.90\ to\ 12.05$
  - **95%** CI for Mean: $CI = 10.98 \pm (\mathbf{1.96} * 0.65) = 9.70\ to\ 12.25$
  - **99%** CI for Mean: $CI = 10.98 \pm (\mathbf{2.58} * 0.65) = 9.30\ to\ 12.66$

For reporting CI: "lower bound" to "upper bound"

- **CI** = interval that should **contain** the population mean $\boldsymbol{\mu}$ in that **% of the samples** (as did occur in these CIs)

Using SE of the Mean to Compare the Sample Mean $\bar{y}$ to an Expected Population Mean $\boldsymbol{\mu}$

- Besides using the SE of the mean to construct a confidence interval around the sample mean $\bar{y}$, we can also use the SE to **compare $\bar{y}$ to an expected population mean $\boldsymbol{\mu}$**

- If we use the **standard normal distribution**, this is known as a "**one-sample z-test**": $\boldsymbol{z} = \frac{\bar{y}-\mu}{SE}$, **where z is a "test statistic"**
  - This test locates $\bar{y}$ onto a new "z" standardized distribution: with $M_Z = 0$ (deviation of $\bar{y}$ from μ) and $SD_Z = 1$ (using **SE** of mean)
  - Our example → expected: $\mu = 10$; sample: $\bar{y} = 10.98$, $SE = 0.65$
    - Is a sample mean = 10.98 *really that different* from expected $\mu = 10$?
    - $z = \frac{\bar{y}-\mu}{SE} = \frac{10.98-10}{0.65} = 1.50$… **ok, so what does $z = 1.50$ actually mean**?

Area Under Standard Normal Curve

Exact probabilities for the area under the curve to the left or right of z can be found by online calculators or statistical software

If $\mu = 10$ was true, we'd find a sample mean $\bar{y} > 10.98$ about 6.68% of the time

Relative to $\boldsymbol{\mu = 10}$, our $\boldsymbol{\bar{y} = 10.98}$ with $\boldsymbol{SE = 0.65}$ puts us here: $\boldsymbol{z = 1.50}$

Said differently, $\boldsymbol{z = 1.50}$ means our $\boldsymbol{\bar{y} = 10.98}$ is $\boldsymbol{+1.50\ SE}$ units away from $\boldsymbol{\mu = 10}$

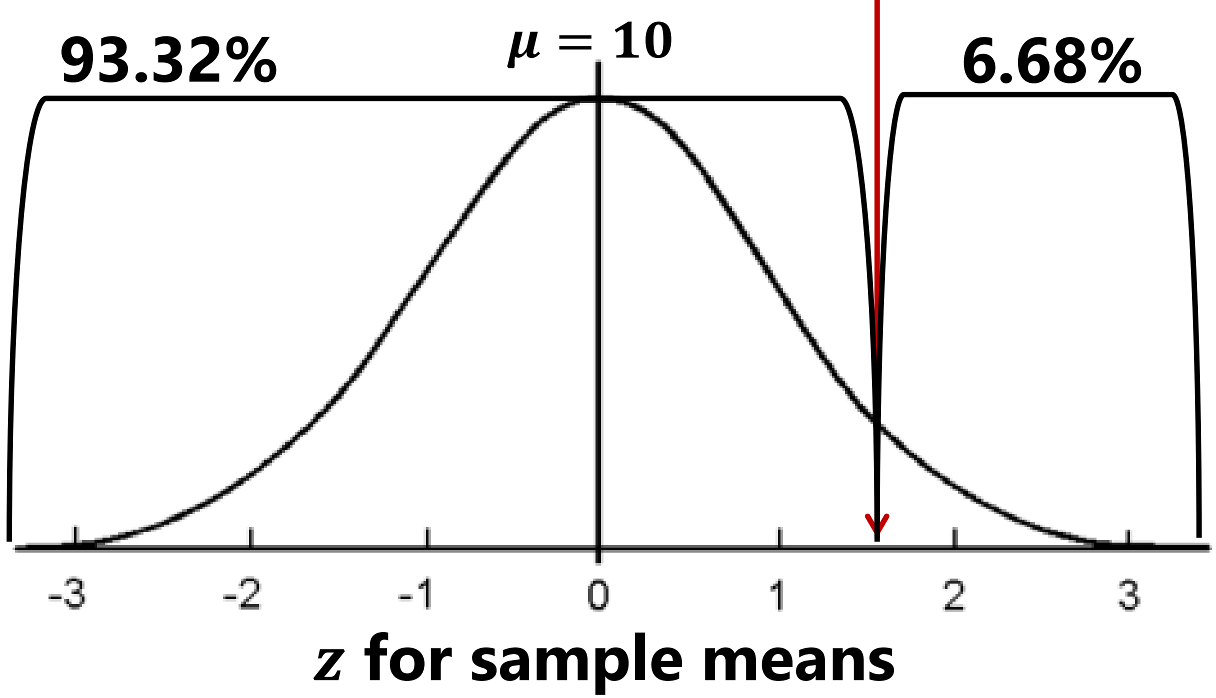


Image adapted from: http://mathcenter.oxford.emory.edu/site/math117/normalDistribution/

So is our sample mean *really that different* from the population mean?

- By **sampling** only some persons from the population, **we expect some fluctuation** in the statistics (e.g., mean and variance) that summarize any one sample, but how different is "**too different**"?
- We **define "too different"** as "only be expected some small percentage of the time" given **three choices made in advance**:
  - What **sampling distribution** characterizes the statistic?
    - For the sample mean, let's stay with **standard normal** distribution (for now)
  - What **percentage** of samples defines "**unexpected**"?
    - This is known as "**alpha level**" and is the **opposite of confidence level**
    - Typically choose **alpha = .05** (or .10 to be lenient, or .01 to be conservative)
  - Is it **possible** to be **unexpected in either direction**?
    - If so, you need a "**two-tailed test**" → allocate alpha % to both sides
    - If not, you need a "**one-tailed test**" → allocate alpha % to one possible side*

More About "Expected" and "Unexpected"

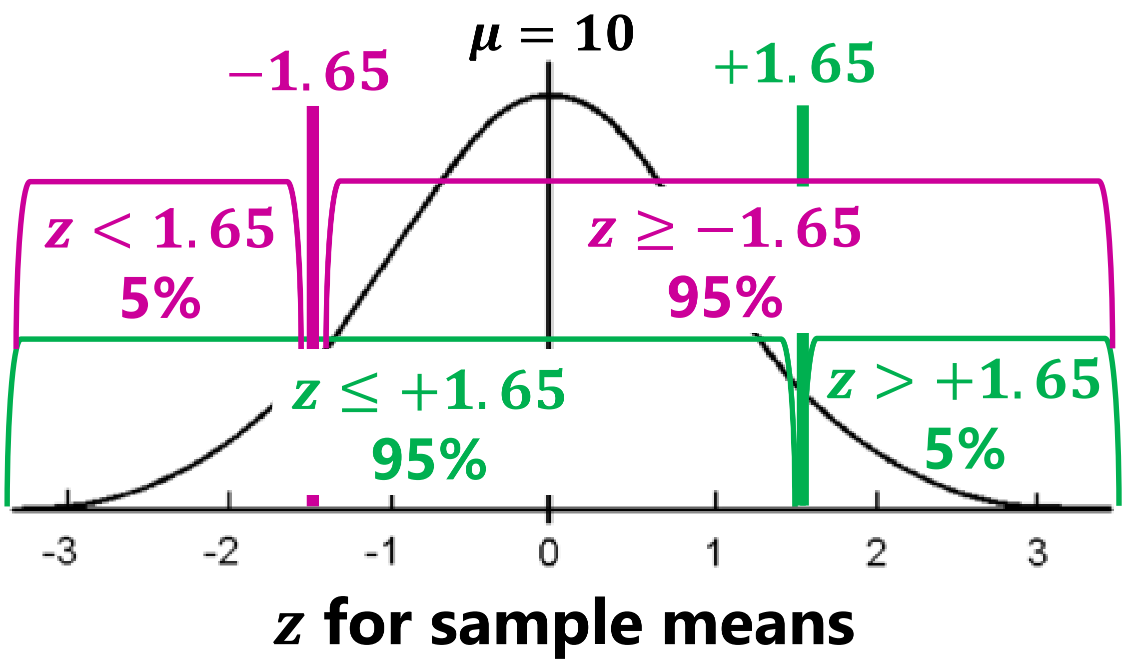
- More generally, this is called a "**Null Hypothesis Significance Test**"; in this example, we are asking "what is the probability of the sample mean $\bar{\boldsymbol{y}}$ if the population mean $\boldsymbol{\mu}$ were true"?
  - A "**hypothesis**" is a statement about a population parameter

- A "**null hypothesis**" ($\boldsymbol{H_0}$) is a statement about the population parameter being equal to some specific (expected) value
  - e.g., in example with sample mean $\bar{y} = 10.98$, $\boldsymbol{H_0}\text{:}\ \boldsymbol{\mu = 10}$

- An "**alternative hypothesis**" ($\boldsymbol{H_A}$) is a statement that contradicts the null hypothesis and **conveys allowed directionality of deviations** from value given by H_0
  - One-tailed test would be $\boldsymbol{H_A\text{:}\ \mu > 10}$ OR $\boldsymbol{H_A\text{:}\ \mu < 10}$
    - Area of unexpected result allocated to one side only
  - Two-tailed tests for "different than": $\boldsymbol{H_A\text{:}\ \mu \neq 10;\ H_A\text{:}\ \mu = !\,10}$
    - Area of unexpected result allocated equally to both sides

Directions of "Unexpected": One-Tailed Tests at Work

- Choices: $\boldsymbol{H_0: \mu = 10}$; probability declared "unexpected" is **alpha = .05** (so **95% "expected"**) → two possible versions one-tailed H_A:
- $\boldsymbol{H_A: \mu > 10}$ → $\boldsymbol{z_{critical} = +1.65}$
  - Tests if $\bar{y}$ is bigger or not bigger
  - If $\bar{y}$ is actually smaller, conclude "not bigger"
- $\boldsymbol{H_A: \mu < 10}$ → $\boldsymbol{z_{critical} = -1.65}$
  - Tests if $\bar{y}$ is smaller or not smaller
  - If $\bar{y}$ actually bigger, conclude "not smaller"



Two-Tailed Test of $\mu \neq 10$: Example Sample of $N = 50$

- **Choices made: at alpha = .05 for a two-tailed test, $z_{critical} = \pm 1.96$**
- Sample statistics: **mean $\bar{y} = 10.98$, SE of mean $= 0.65$**
- 95% CI for Mean: $CI = 10.98 \pm (1.96 * 0.65) = 9.70 \textit{ to } 12.25$ (so has μ)
- One-sample z-test given H_0 that $\mu = 10$: $z = \frac{\bar{y} - \mu}{SE} = \frac{10.98 - 10}{0.65} = 1.50$
- **Exact two-tailed p-value for $z = 1.50$ is $p = 0.1336$**

Two-sided **p-value** = probability of a **more extreme z test statistic** than was found: $6.68 * 2 = 13.36$

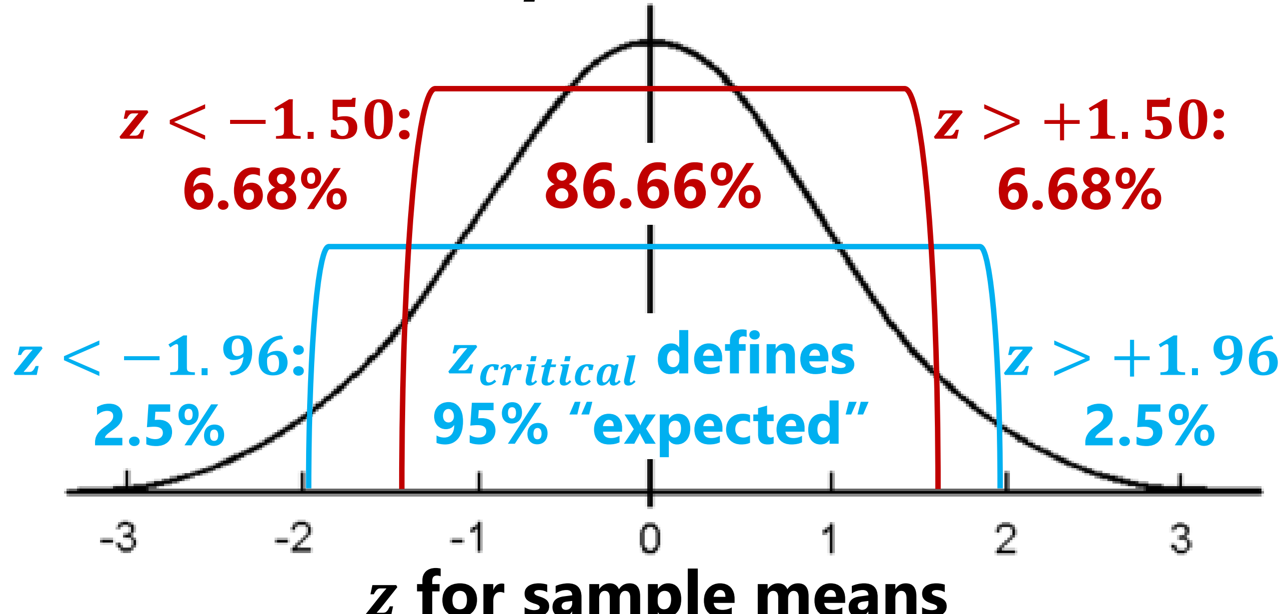


Image adapted from: http://mathcenter.oxford.emory.edu/site/math117/normalDistribution/

Decision Language for Test-Statistics

- Calculation of test-statistics (like z) and their p-values are more informally called "**significance tests**" (against a null hypothesis H_0)

- If the **test-statistic exceeds** the chosen distribution's critical value(s), then the obtained $\boldsymbol{p}$**-value is less than the chosen alpha** level:
  - You "**reject the null hypothesis**": it is sufficiently **unexpected** to get an observed test-statistic that extreme *if the null hypothesis were true*
  - So the **test result** is labeled "**statistically significant**"

- If the **test-statistic does not exceed** the distribution's critical value(s), then the obtained $\boldsymbol{p}$**-value is greater than or equal to the chosen alpha** level:
  - You "**do not reject* the null hypothesis**"—it is sufficiently **expected** to get an observed test-statistic that extreme *if the null hypothesis were true*
    - * You CANNOT SAY "accept the null hypothesis" or you will be chastised!
    - * I think you can say "retain the null hypothesis" but some may quibble on that
  - So the **test result** is labeled "statistically **<u>non</u>significant***"
    - * Do not say "insignificant" because that is a value judgment—instead say "not significant" or "nonsignificant" (conventionally written as one word)

Decision Language for Example Sample

- **Choices made: use standard normal, at alpha = .05 for a two-tailed test, $z_{critical} = \pm 1.96$, population mean expected to be $\mu = 10$**
- Obtained z **test-statistic** and p**-value**: $z = 1.50$, $p = 0.1336$
- If H_0 were true (if $\mu = 10$), we would see a sample mean of $\bar{y} = 10.98$ (**SE of mean** $= 0.65$) that is more than 1.5 standard deviations away from the mean (either **too high or too low**, so beyond $\pm z = 1.50$) approximately **13.36%** of the time (6.68% $z > 1.50$; 6.68% $z < -1.50$)
- Because $z > 1.96$ and so $p > .05$, **the test result is nonsignificant**: $\bar{y} = 10.98$ is **nonsignificantly greater than expected** $\mu = 10$

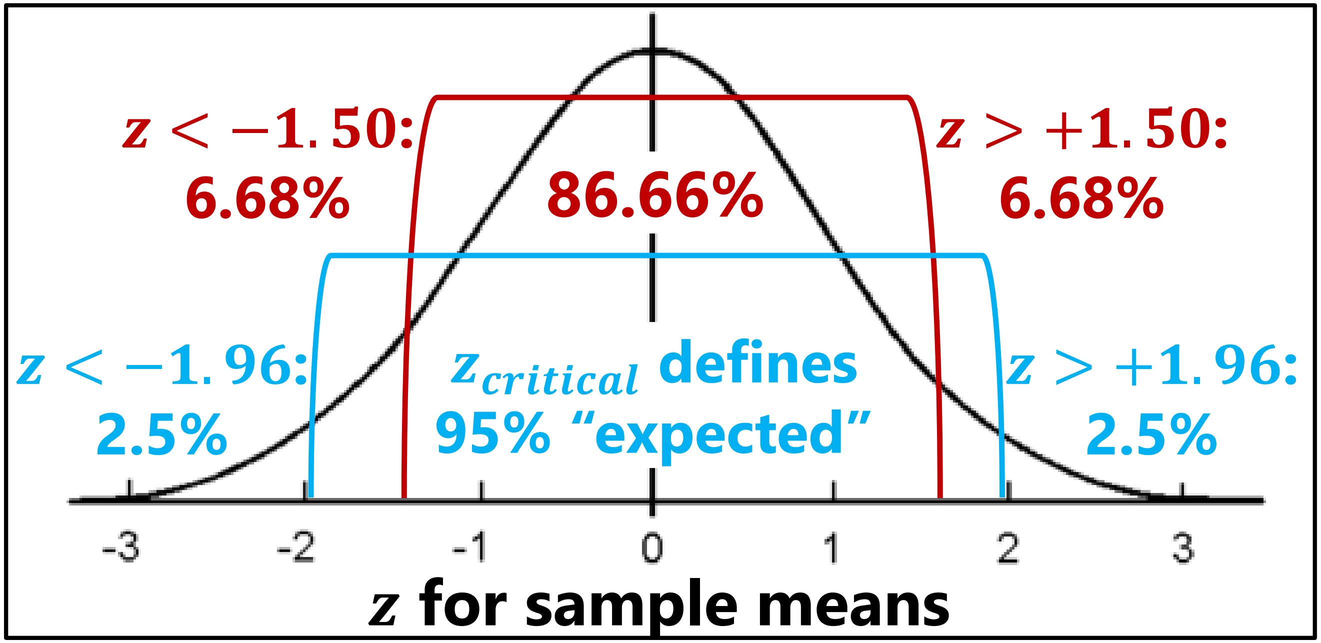


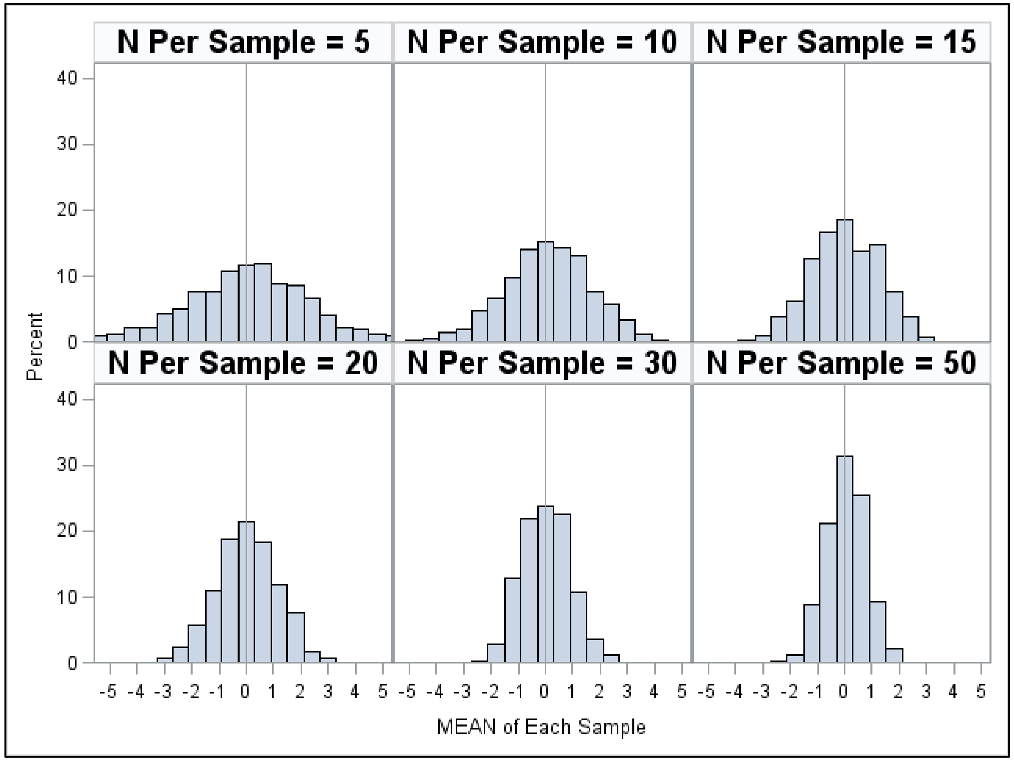
Image adapted from: http://mathcenter.oxford.emory.edu/site/math117/normalDistribution/

Using the SE of the Sample Mean to Make Inferences to the Population Mean

- So far we've seen **two inferential uses** of the **SE** of the mean:
  - To create a **confidence interval**: limits of the range that should **contain** the population mean $\boldsymbol{\mu}$ in chosen **% of the samples**
  - To create a **test statistic** how obtained sample mean $\bar{\boldsymbol{y}}$ differs from an expected population mean $\boldsymbol{\mu}$ (where $\boldsymbol{\mu}$ is the null hypothesis, H_0)
    - If $\boldsymbol{\mu}$ is true, how often would we find a more extreme value of $\bar{\boldsymbol{y}}$?
- We've seen both uses require **three choices made in advance:**
  - Where "**unexpected**" begins: expressed as either confidence level (e.g., 95% expected) or **alpha level** (e.g., 5% unexpected)
  - **Direction** of unexpected: Either too high or too low → **two-tailed**
  - **Which PDF** describes the statistic's sampling distribution—provides the **critical values** to map your **% unexpected** onto your sample
    - In real data we don't know for sure which distribution is "correct", but let's see how the **normal distribution** worked in our simulated data…

95% CIs for the Mean via Normal Distribution

1000 samples drawn for each N from y_i : Mean $\mu = 10$, SD $\sigma = 5$



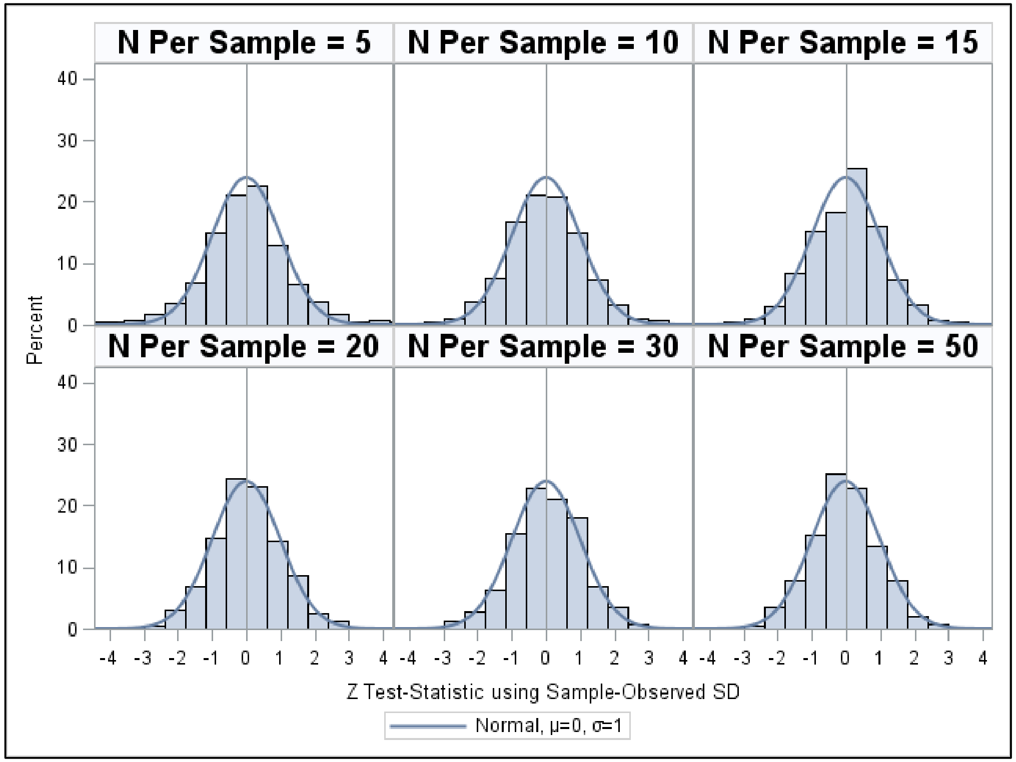
N	% of CIs with $\mu = 0$	Mean SE with: σ	Mean SE with: s
5	**88.3**	2.24	2.13
10	**90.8**	1.58	1.55
15	**93.1**	1.29	1.28
20	**94.3**	1.12	1.11
30	**94.4**	0.91	0.91
50	**94.1**	0.71	0.71

Should be 95%!

The **95% CI** for a sample mean provides the interval that should contain the population mean in 95% of the samples. But in reality, only **88–94%** of CIs for these samples contained the population mean. **So what happened ????**

Test $\bar{y}$ against μ via Normal Distribution

1000 samples drawn for each N from y_i : Mean $\mu = 10$, SD $\sigma = 5$



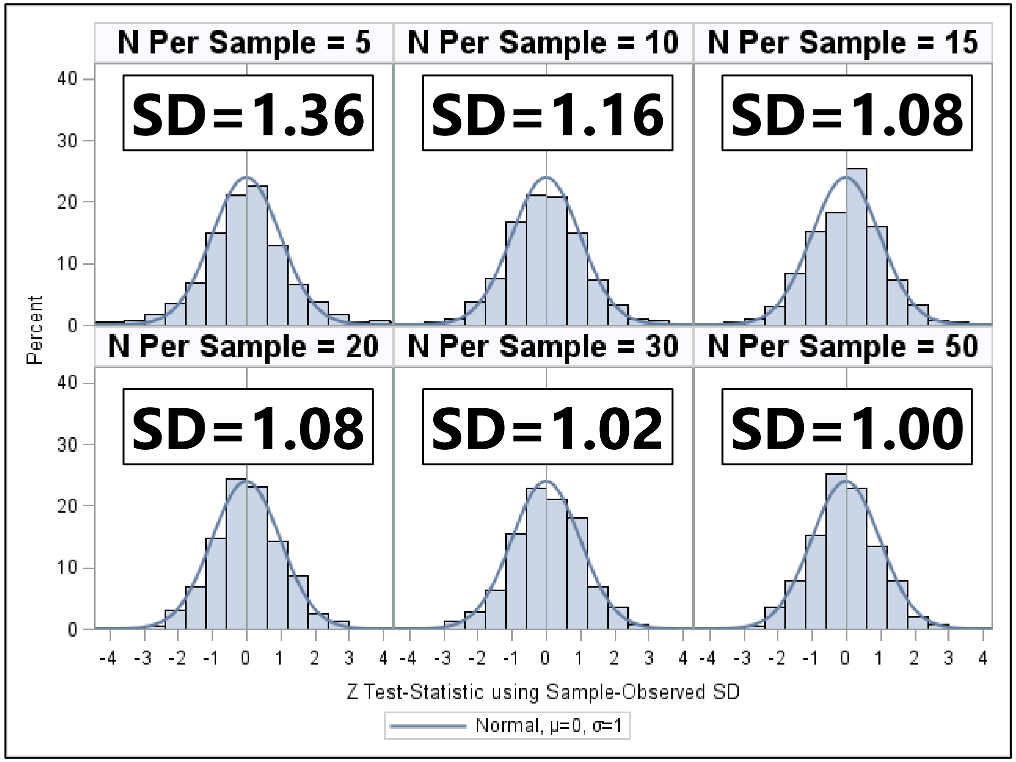
N	% of tests with $p < .05$	Mean SE with: σ	Mean SE with: s
5	**11.7**	2.24	2.13
10	**9.2**	1.58	1.55
15	**6.9**	1.29	1.28
20	**5.7**	1.12	1.11
30	**5.6**	0.91	0.91
50	**5.9**	0.71	0.71

Should be 5%!

If the standardized normal distribution accurately characterized the sampling distribution of the mean, then we would have z test-statistics more extreme than the chosen critical value of ±1.96 **less than 5% of the time**. But in reality, **up to 11.7%** of these z test-statistics were found to be "significant". **So what happened ????**

Test $\bar{y}$ against μ via Normal Distribution: s

1000 samples drawn for each N from y_i : Mean $\mu = 10$, SD $\sigma = 5$



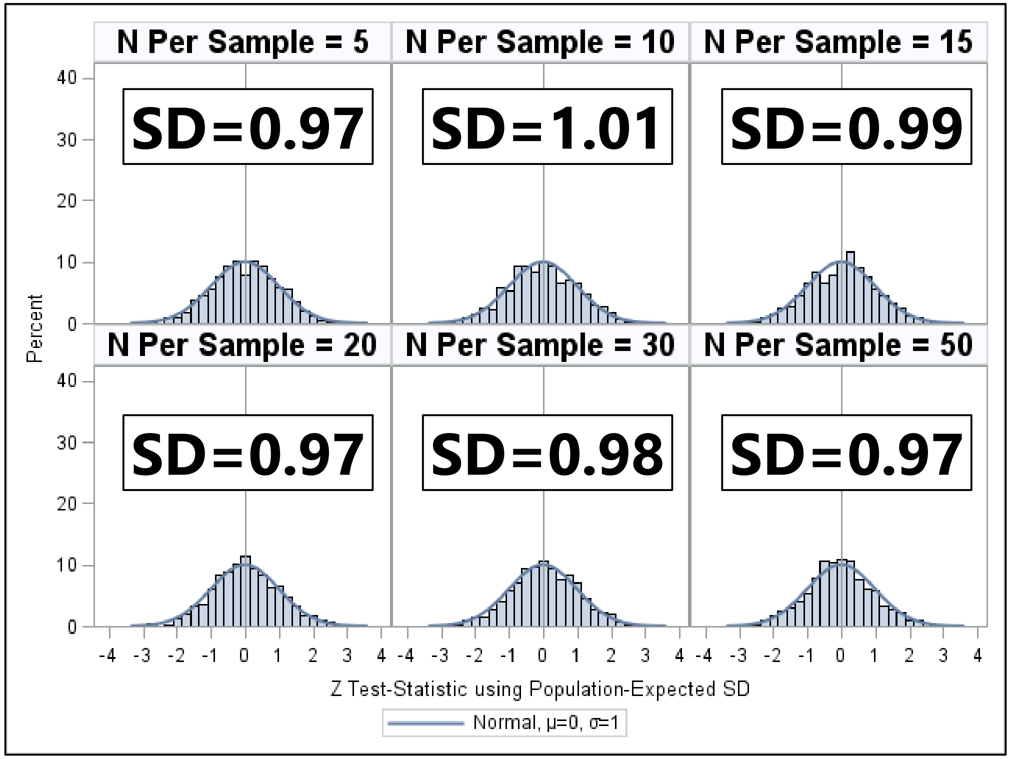
N	% of tests with $p < .05$	Mean SE with:	
		σ	s
5	**11.7**	2.24	**2.13**
10	**9.2**	1.58	**1.55**
15	**6.9**	1.29	**1.28**
20	**5.7**	1.12	**1.11**
30	**5.6**	0.91	**0.91**
50	**5.9**	0.71	**0.71**

Should be 5%!

The **SD** for each of these z **test-statistics** was supposed to be **1.00** (to match the standard normal distribution), but the **observed SDs were larger as N decreased**. This is partially because the observed sample SD (s) was used in computing the SE instead of the expected population SD (σ). What would happen if we used σ instead?

Test $\bar{y}$ against μ via Normal Distribution: σ

1000 samples drawn for each N from y_i : Mean $\mu = 10$, SD $\sigma = 5$



N	% of tests with $p < .05$	Mean SE with: σ	Mean SE with: s
5	**4.1**	**2.24**	2.13
10	**5.0**	**1.58**	1.55
15	**4.6**	**1.29**	1.28
20	**4.1**	**1.12**	1.11
30	**4.9**	**0.91**	0.91
50	**4.7**	**0.71**	0.71

Closer to 5% ☺

After switching from the observed sample SD (s) to the expected population SD (σ) in computing the SE, the **SD** for each of these z **test-statistics** is closer to the **1.00** it should be, and about **5%** of these z test-statistics were flagged as "unexpected" as they should be. **But what if you don't know σ??? Beer to the rescue! No, really…**

What Went Wrong? Beer to the Rescue!

- As we just saw, the standard normal doesn't fit well in small samples
- True story: this discovery is credited to William S. Gosset, who began working for Guinness Brewery in 1899 testing batches of hops for acceptability relative to a target population mean
  - Because his testing could take a whole day and it could take a full year to grow a crop, his sample sizes were tiny (like 3-4 batches in a sample)
  - He computed *z* test-statistics for each sample, and those whose mean was deemed outside the target mean ("unexpected") then had further testing
    - **But 3 times more than expected, the samples were actually ok… huh**?
    - So Guinness let him go get a graduate degree in statistics to try to figure out why, and he did so by hand: He drew 750 samples of N=4 by shuffling 3000 cards (whose population mean he knew), and derived a new distribution
    - Guinness prohibited employees from publishing anything (i.e., trade secrets), but Gosset convinced them to let him publish his finding as author "Student"
    - And "**student's *t***" was born! Let's compare **standard normal** and ***t*** distributions…

Source: DeVeaux, R. D., Velleman, P. F., & Bock, D. E. (2004 p. 476-477). *Intro Stats* (4th ed). Boston: Pearson.

z Standard Normal ignores N…

Metric: $\boldsymbol{M_Z = 0}$ measures deviations of $\boldsymbol{\bar{y}}$ from $\boldsymbol{\mu}$ in **SE** units (as $\boldsymbol{SD_Z = 1}$)

Confidence Intervals using standard normal distribution have these **z "critical" values:**

90% within $z = \pm 1.65$
95% within $z = \pm 1.96$
99% within $z = \pm 2.58$

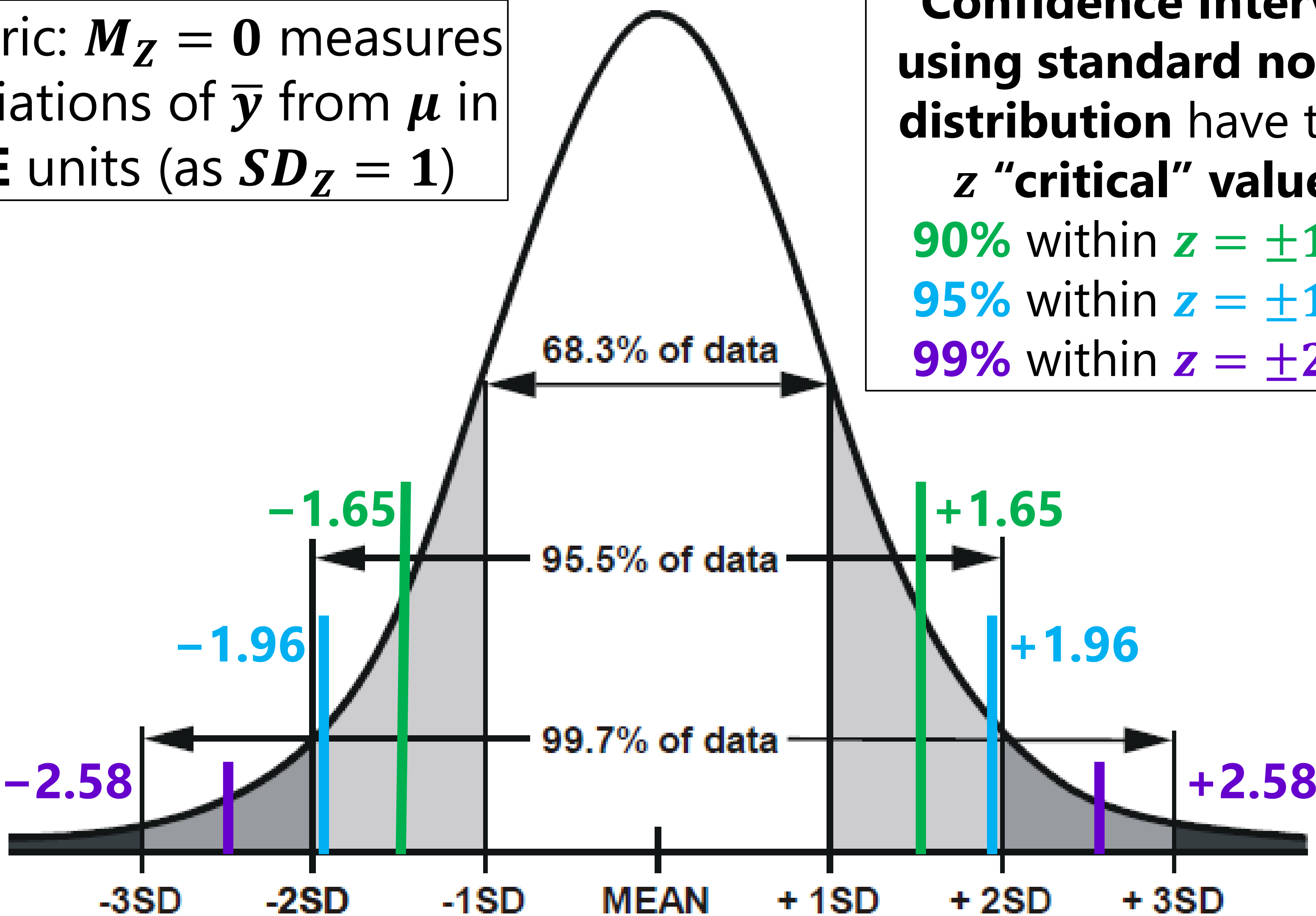
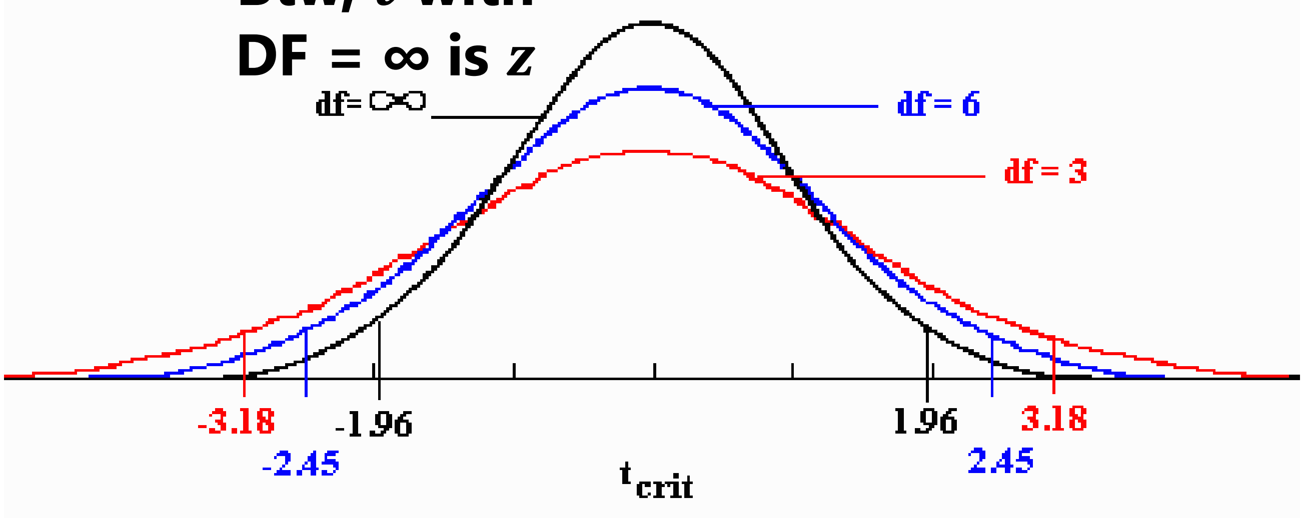


Image adapted from: http://my.ilstu.edu/~gjin/hsc204-hed/Module-5-Summary-Measure-2/Module-5-Summary-Measure-28.html

Meet **Student's** t Distribution: Where Sample Size N Matters!

- Both $\boldsymbol{z}$ (standard normal) and $\boldsymbol{t}$ distributions have the same metric: $\boldsymbol{M=0}$, $\boldsymbol{SD=1}$ (to translate $\bar{y} \rightarrow \mu$ given **SD → SE of the mean**)
- But $\boldsymbol{t}$ is flatter than $\boldsymbol{z}$, more so with **fewer "denominator degrees of freedom": DF** $= \boldsymbol{N-1}$ (for now; stay tuned)

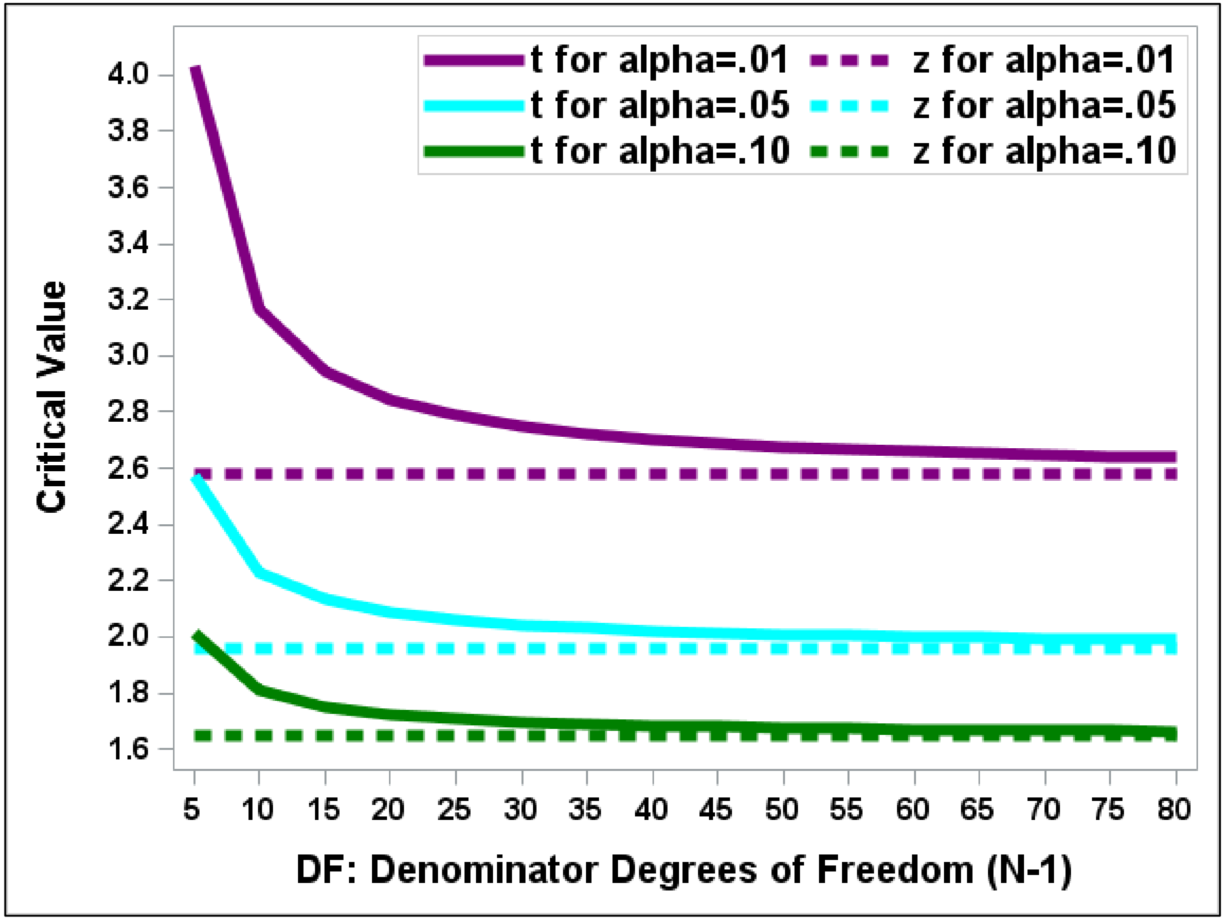
Btw, $\boldsymbol{t}$ **with DF = ∞ is** $\boldsymbol{z}$



- $\boldsymbol{t_{critical}}$ values for **alpha = .05 by DF** shown here
- **With smaller** $\boldsymbol{N}$, have to go farther out to **get to 5%**

Image borrowed from: http://faculty.cbu.ca/~erudiuk/IntroBook/sbk24m.htm

Critical Values for t versus z Distributions



With smaller N (fewer DF), greater t test-statistics are needed to declare $\bar{y}$ as "unexpectedly different" from μ (i.e., to cross the alpha threshold to be "**significant**")

z doesn't use DF

In the olden days, one needed to refer to tables of $t_{critical}$ values for a given alpha and DF, but now statistical software can give you the **exact p-value**: the probability of a more extreme t test-statistic than you found if the null hypothesis H_0 were true

Using t Distribution Instead of z: Example Sample of $N = 50$

- **Choices made: at two-tailed alpha = .05, $t_{critical} = \pm 2.01$ (not $z = 1.96$)**
- Sample statistics: **mean $\bar{y} = 10.98$, SE of mean $= 0.65$**
- 95% CI for Mean: $CI = 10.98 \pm (2.01 * 0.65) = 9.67 \text{ to } 12.28$ (so has μ)
- One-sample t-test **given H_0 that $\mu = 10$**: $t = \frac{\bar{y}-\mu}{SE} = \frac{10.98-10}{0.65} = 1.50$
- **Exact p-value for $t = 1.50$ is $p = 0.1396$**

Two-sided **p-value** = probability of a **more extreme z test statistic** than was found: $6.98 * 2 = 13.96$

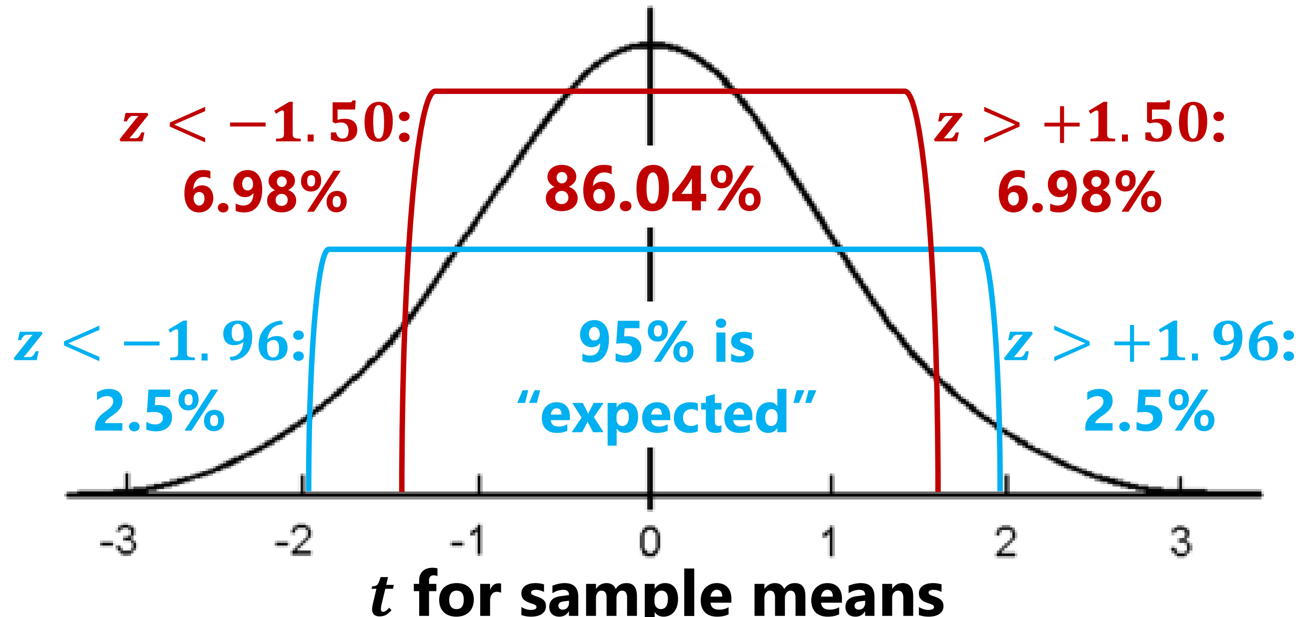


Image adapted from: http://mathcenter.oxford.emory.edu/site/math117/normalDistribution/

Test $\bar{y}$ against μ via t Distribution instead of z

1000 samples drawn for each N from y_i : Mean $\mu = 10$, SD $\sigma = 5$

N	% of tests with $p < .05$		% of CIs with $\mu = 0$		Mean SE with:	
	z	t	z	t	σ	s
5	11.7	**5.3**	88.3	**94.7**	2.24	2.13
10	9.2	**4.9**	90.8	**95.1**	1.58	1.55
15	6.9	**4.9**	93.1	**95.1**	1.29	1.28
20	5.7	**4.1**	94.3	**95.9**	1.12	1.11
30	5.6	**5.0**	94.4	**95.0**	0.91	0.91
50	5.9	**4.8**	94.1	**95.2**	0.71	0.71

Using the t **distribution**, which takes into account **denominator degrees of freedom**, resulted in confidence intervals (CIs) that contained the population mean μ closer to the chosen 95%, or equivalently, 5% of tests that found the difference between the sample mean $\bar{y}$ and expected μ to be "significant"

More About "Degrees of Freedom"

- More specifically, we are focusing for now on the **denominator term** in the formulas for estimating the mean (N) and variance ($N-1$)
  - "Estimate" means "find the best value"; for now we can use off-the-shelf formulas
  - This term is known more generally as **denominator degrees of freedom**, abbreviated as DF_{den} (or DDF); it is referred to as just "DF" with t test-statistics
- DF_{den} is based on the concept that to fully describe all values in a variable, we could **compute up to N statistics**—this is our starting point for DF_{den}
- For each statistic we estimate to describe a variable, we "spend" 1 DF_{den} and reduce the denominator term accordingly to reflect the remainder
    - Real-world analogy: Weight Watchers "points" (see also 1980's "Deal-A-Meal")
  - For example: mean $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$, variance $s^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$
    - DF_{den} for the mean starts out at N, because we haven't already computed anything that is needed in order to estimate the mean
    - But DF_{den} for the variance is $N-1$ to account for having already spent 1 DF_{den} to estimate $\bar{y}$ for use in estimating the variance
  - This trend will continue as we estimate other statistics within models…

Summary: Inference for Sample Means of Quantitative Variables

- **SE of the mean** indexes mean's **inconsistency** across samples (is a proxy for SD of the sampling distribution for $\bar{\boldsymbol{y}}_{\boldsymbol{s}}$)
  - $SE = \frac{\boldsymbol{\sigma}}{\sqrt{N}}$ if population SD is known; $SE = \frac{\boldsymbol{s}}{\sqrt{N}}$ if using sample SD
  - The means of samples with **more within-sample variance** and **smaller sample sizes** have **larger SEs** (i.e., more imprecision)

- SE is used to create confidence intervals (range expected to contain the population mean $\boldsymbol{\mu}$ in that % of samples) and/or to form a test-statistic that compares sample $\bar{\boldsymbol{y}}$ to expected $\boldsymbol{\mu}$
  - Safest strategy is to use a ***t*-distribution with denominator DF** ($\boldsymbol{DF} = N - 1$ here) to get critical value for CI and/or exact p-value
    - $CI = Estimate \pm (\boldsymbol{t}_{\boldsymbol{critical}} * SE)$; one-sample "$\boldsymbol{t}$-test": $\boldsymbol{t} = \frac{\bar{\boldsymbol{y}} - \boldsymbol{\mu}}{\boldsymbol{SE}}$
  - If population SD is known or N is "big enough", standard normal → ok
    - $CI = Estimate \pm (\boldsymbol{z}_{\boldsymbol{critical}} * SE)$; one-sample "$\boldsymbol{z}$-test": $\boldsymbol{z} = \frac{\bar{\boldsymbol{y}} - \boldsymbol{\mu}}{\boldsymbol{SE}}$

Real Example: Twinning Effect*

- **Twinning Effect**: Developmental delay in twins relative to singletons
- Demonstrated if 95% CI for sample $\bar{y}$ was **below** expected $\mu = 100$

From Table 1 of Rice et al. (2018)

Phenotype	Age	Zygosity	*n*	Mean	*SE*	Lower CI	Upper CI	CI excludes population mean
PPVT-3 Vocabulary	4	DZ	771	96.78	0.59	95.62	97.93	–
		MZ	357	93.91	0.91	92.11	95.70	–
	6	DZ	798	101.93	0.45	101.06	102.81	+
		MZ	372	100.28	0.79	98.73	101.83	

- $CI = Estimate \pm (critical * SE)$; for $DF = 770$, $t_{critical}$ ~1.96
  - **For DZ age 4**, $\mathbf{95\%}$ $\boldsymbol{CI = 96.78 \pm (1.96 * 0.59) = 95.62\ to\ 97.93}$
  - Because the interval is below $\mu = 100$, there is evidence of a twinning effect: significantly lower sample mean than expected ($\mu = 100$ → standardized test)
- "**one-sample *t*-test**": $t = \frac{\bar{y}-\mu}{SE} = \frac{96.78-100}{0.59} = -5.48$, two-tailed $p < .0001$
  - If $\boldsymbol{\mu} = 100$, $\bar{\boldsymbol{y}} = 96.78$ (5+ SDs from the mean) would be found < 0.01% of time

* Source: Rice, M. L., Zubrick, S. R., Taylor, C. L., Hoffman, L., & Gayán, J. (2018). Longitudinal study of language and speech of twins at 4 and 6 years: Twinning effects decrease, zygosity effects disappear, and heritability increases. *Journal of Speech-Language-Hearing Research, 61*(1), 79-93. Note: SE was from a multilevel model accounting for twin dependency.

Real Example: Twinning Effect*

- **Twinning Effect**: Developmental delay in twins relative to singletons
- Demonstrated if 95% CI for sample $\bar{y}$ was **below** expected $\mu = 100$

PPVT from Table 1 of Rice et al. (2018)

Age	Zygosity	*n*	Mean	*SE*	Lower CI	Upper CI	CI excludes population mean
4	DZ	771	96.78	0.59	95.62	97.93	–
	MZ	357	93.91	0.91	92.11	95.70	–
6	DZ	798	101.93	0.45	101.06	102.81	+
	MZ	372	100.28	0.79	98.73	101.83	

t-value	2-tailed p-value
-5.48	$< .0001$
-6.69	$< .0001$
4.29	$< .0001$
0.35	$= .7232$

- **Age 4** shows evidence of **significant** twinning effect: if $\boldsymbol{\mu} = 100$, $\bar{y}$ estimates as extreme as these would be found < 0.01% of the time
- **Age 6 DZ** result is also **significant**, but in the **opposite direction**
  - If we had used a one-tailed test for $\boldsymbol{\mu} < 100$, we would say the result is nonsignificant ($\bar{y}$ was not < 100), but that would mis-state the real story!
- **Age 6 MZ** result is **nonsignificant**: more extreme expected 72.32% of time

* Source: Rice, M. L., Zubrick, S. R., Taylor, C. L., Hoffman, L., & Gayán, J. (2018). Longitudinal study of language and speech of twins at 4 and 6 years: Twinning effects decrease, zygosity effects disappear, and heritability increases. *Journal of Speech-Language-Hearing Research, 61*(1), 79-93. Note: SE was from a multilevel model accounting for twin dependency.

Example One-Sample t-Test in **SAS**: Is the mean years of education different than 12?

```
* TTEST to compare sample mean to H0=expected at alpha=.05;
* CI=equal also requests confidence interval for SD;
PROC TTEST DATA=work.Example1 H0=12 SIDES=2 ALPHA=.05
           CI=EQUAL PLOTS=NONE;
VAR educ; RUN;
```

N	Mean	Std Dev	Std Err	Minimum	Maximum
734	13.8120	2.9093	**0.1074**	2.0000	20.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
13.8120	**13.6012**	**14.0228**	2.9093	2.7677	3.0663

DF	t Value	Pr > |t|
733	**16.87**	**<.0001**

- $\bar{y} = 13.812, \mathbf{SE} = 0.1074$
  - **95%** $\boldsymbol{CI} = 13.601$ to 14.023
- $\boldsymbol{t} = \frac{13.81-12}{0.11} = \mathbf{16.87},\ \boldsymbol{p} < \mathbf{.0001}$
  - If the true population mean was $\boldsymbol{\mu} = \mathbf{12}$ **years**, a more extreme sample mean than $\bar{y} = \mathbf{13.81}$ (± 16.87 SDs away) would be **found < 0.01% of the time**
  - 13.81 is greater than 12 (<u>significantly</u> because $p < .05$)

Example One-Sample *t*-test in **STATA**: Is the mean years of education different than **12**?

```
// TTEST to compare sample mean to H0=expected at alpha=.05
ttest educ==12, level(95)
```

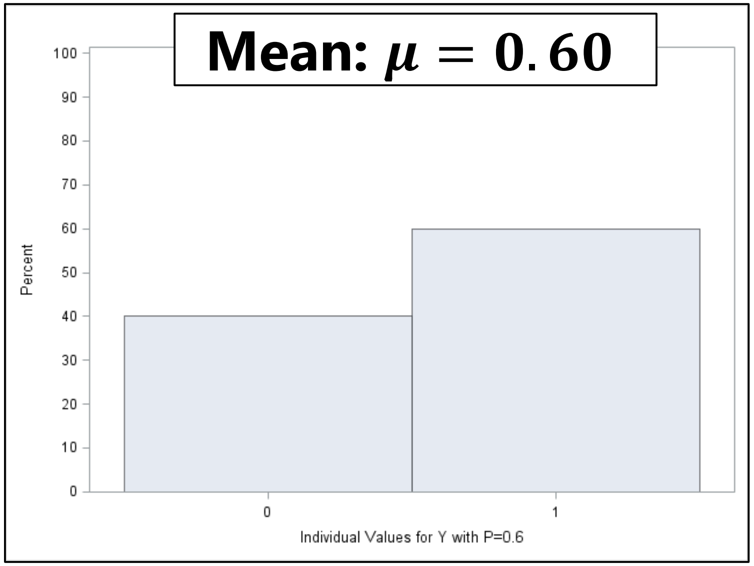
```
.    ttest educ==12, level(95)
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    educ |     734    13.81199    .1073836    2.909282    13.60117    14.02281
------------------------------------------------------------------------------
    mean = mean(educ)                                             t =  16.8740
Ho: mean = 12                                    degrees of freedom =      733

    Ha: mean < 12                 Ha: mean != 12                 Ha: mean > 12
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

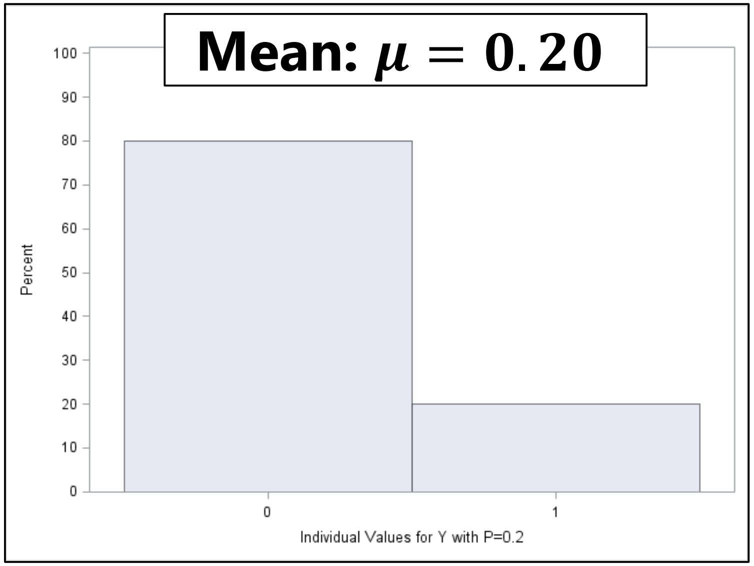
```
// CI requests confidence interval for variance separately
ci variances educ, level(95)
```

```
.    ci variances educ, level(95)
    Variable |        Obs       Variance       [95% Conf. Interval]
-------------+------------------------------------------------------
        educ |        734       8.463922       7.660085    9.401962
```

What about Means of Binary Variables?

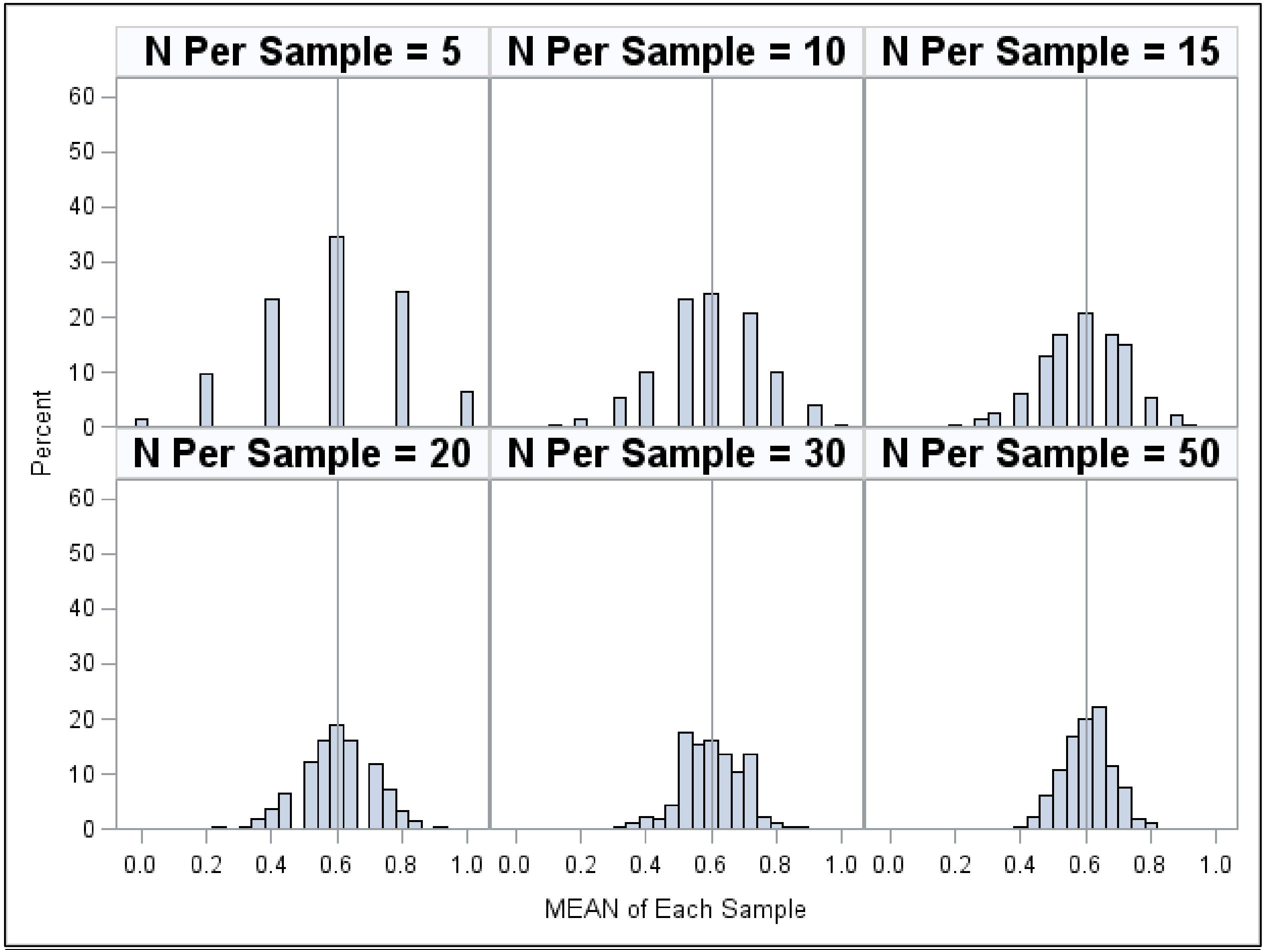
- In **binary** variables (**0 or 1** values), $\bar{y}$ is labeled as $\boldsymbol{p}$, the proportion of **1** values (and $\boldsymbol{q}$ is the proportion of **0** values)
- $\boldsymbol{N}$ **and** $\boldsymbol{p}$ **affect how close** $\bar{y}$ **is to true** $\boldsymbol{\mu}$ (because $VAR = \boldsymbol{p} * \boldsymbol{q}$)
  - 2 fake binary variables* for a population of 100,00 fake people
  - Mean: $\boldsymbol{\mu = 0.60}$ or $\boldsymbol{0.20}$, so VAR: $\boldsymbol{\sigma^2 = 0.24}$ or $\boldsymbol{0.16}$





* Used a “Bernoulli” distribution here to generate $\boldsymbol{y_i}$ (two categories)

1000 samples each for different N…

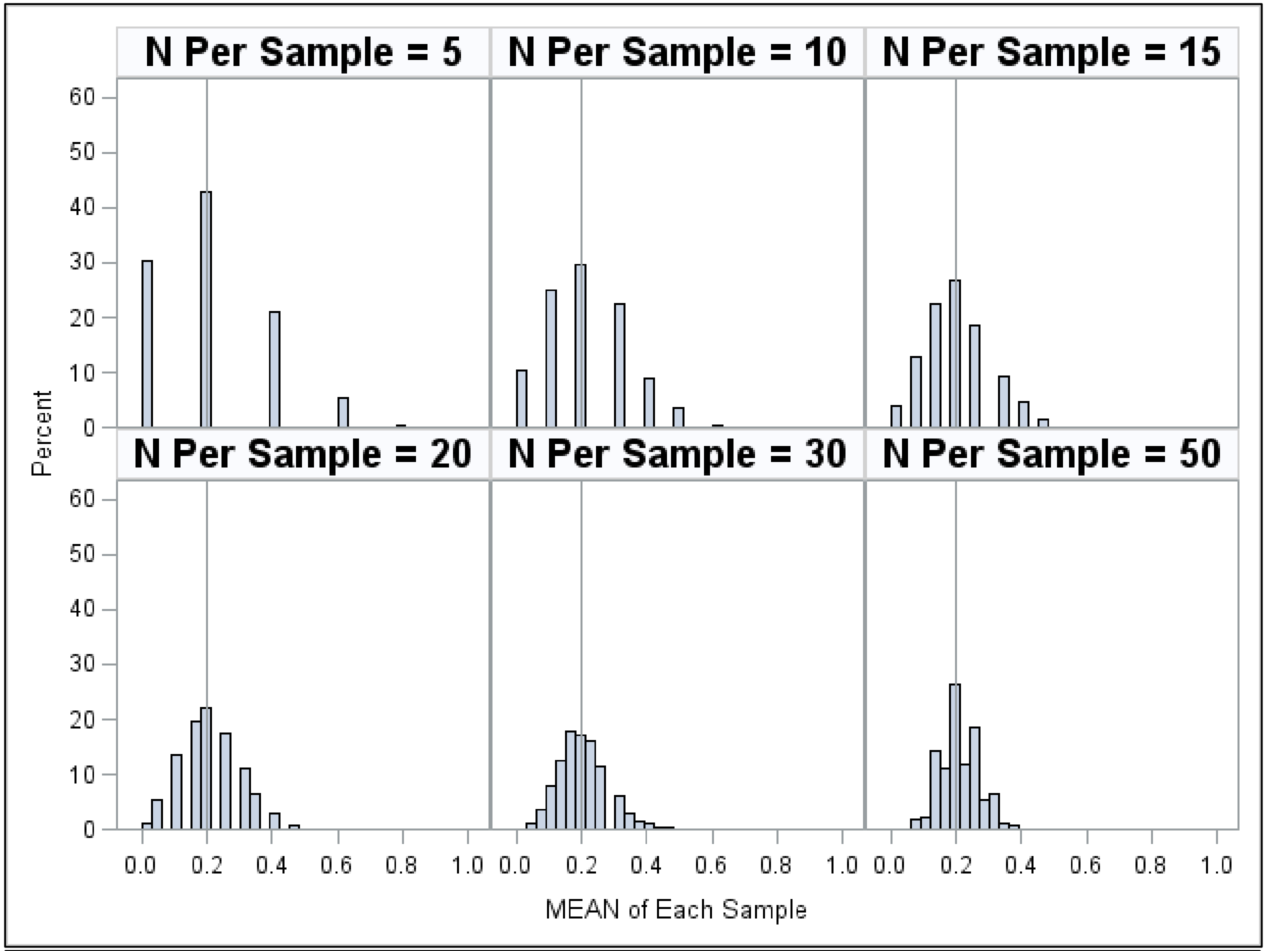


The sample mean of binary variables ($\bar{y}$, called "p") follows a **binomial distribution** (using only N and μ) that can be approximated by a normal distribution in larger samples

- Population values:
  Mean $\mu = 0.60$
  (so VAR $\sigma^2 = 0.24$)
- **More N → less SD in $\bar{y}_s$** across samples

N Per Sample	Mean $\bar{y}_s$	SD $\bar{y}_s$
5	0.58	0.22
10	0.59	0.16
15	0.60	0.13
20	0.60	0.11
30	0.60	0.09
50	0.60	0.07

1000 samples each for different N…



The sample mean of binary variables ($\bar{y}$, called "p") follows **binomial distribution** (using only N and μ) that becomes more skewed the farther away μ is from the midpoint **0.5**

- Population values:
  Mean $\mu = 0.20$
  (so VAR $\sigma^2 = 0.16$)
- **More N → less SD in $\bar{y}_s$** across samples

N Per Sample	Mean $\bar{y}_s$	SD $\bar{y}_s$
5	0.21	0.18
10	0.21	0.13
15	0.20	0.10
20	0.20	0.09
30	0.20	0.07
50	0.20	0.06

Inference for Means of Binary Variables

- The same issues with inference about the mean of quantitative variables occur for the mean of binary variables ($\bar{y}$, called the proportion p)
- **Two conditions** should be met to use z **standard normal approximation** to binomial distribution: $Np > 5$ and $Nq > 5$ (or > 10 in some sources)
  - In Example 1, I used this normal approximation to ensure consistent results across SAS and STATA
- Otherwise, **numerous (non-t) "fixes"** have been proposed that:
  - Ensure CI for proportion p stays within boundaries of 0 and 1 (CI may need to be asymmetric as a result)
  - Account for more inconsistency with smaller N and extreme p
  - Include various "continuity corrections" and "exact statistics" that may involve resampling techniques to derive an empirical SE
- For more details in software implementation:
  - SAS PROC FREQ documentation
  - STATA exact tests

For more info, ask The Google about "categorical data"

Example One-Sample Proportion Test in **SAS**: Is the proportion of HS non-graduates different than **.10**?

```
* FREQ: compare proportion to BINOMIAL P=expected at alpha=.05;
* Specify LEVEL= to test proportion of 1 values against H0;
* CL requests confidence interval for proportion;
PROC FREQ DATA=work.Example1;
TABLE lessHS / CL BINOMIAL(LEVEL="1" P=.10) ALPHA=.05;
RUN;
```

lessHS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	643	87.60	643	87.60
1	91	**12.40**	734	100.00

Binomial Proportion lessHS = 1	
Proportion	**0.1240**
ASE	**0.0122**
95% Lower Conf Limit	**0.1001**
95% Upper Conf Limit	**0.1478**
Exact Conf Limits	
95% Lower Conf Limit	0.1010
95% Upper Conf Limit	0.1500

Test of H0: Proportion = 0.1	
ASE under H0	0.0111
Z	**2.1654**
One-sided Pr > Z	0.0152
Two-sided Pr > |Z|	0.0304

SE for z-test uses σ instead of s

- $\bar{y} = 0.124, \mathbf{SE} = 0.0122$ (using s)
  - $95\%\ CI = 0.1001$ to 0.1478
- $z = \frac{0.124-0.10}{0.0111} = \mathbf{2.165},\ p = \mathbf{.0304}$
  - If $\mu = \mathbf{0.10}$, a more extreme sample mean than $\bar{y} = \mathbf{0.124}$ (± 2.165 SDs away) would be found ~ **3.04%** of the time
  - 0.124 is greater than 0.10 (significantly because $p < .05$)

Example One-Sample Proportion Test in **STATA**:
Is the proportion of HS non-graduates different than **.10**?

```
// PRTEST: compare sample proportion to variable=expected at alpha=.05
// STATA always tests proportion of 1 values against H0
prtest lessHS==.10, level(95)

One-sample test of proportion                  lessHS: Number of obs =      734
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lessHS |   .1239782   .0121642                      .1001369    .1478195
------------------------------------------------------------------------------
    p = proportion(lessHS)                                        z =   2.1654
Ho: p = 0.1

     Ha: p < 0.1                 Ha: p != 0.1                 Ha: p > 0.1
 Pr(Z < z) = 0.9848         Pr(|Z| > |z|) = 0.0304          Pr(Z > z) = 0.0152
```

- $\bar{y} = 0.124, \mathbf{SE} = 0.0122$ (using s); $\mathbf{95\%}\ \boldsymbol{CI} = 0.1001$ to 0.1478
- $z = \frac{0.124-0.10}{0.0111} = \mathbf{2.165},\ \boldsymbol{p} = \mathbf{.0304}$
  - If $\mu = \mathbf{0.10}$, a more extreme sample mean than $\bar{y} = \mathbf{0.124}$ (± 2.165 SDs away) would be found ~ **3.04%** of the time
  - 0.124 is greater than 0.10 (significantly because $p < .05$)

Opposite One-Sample Proportion Test in **SAS**:

Is the proportion of HS graduates different than **.90**?

```
* FREQ: compare proportion to BINOMIAL P=expected at alpha=.05;
* Specify LEVEL= to test proportion of 1 values against H0;
* CL requests confidence interval for proportion;
PROC FREQ DATA=work.Example1;
TABLE gradHS / CL BINOMIAL(LEVEL="1" P=.90) ALPHA=.05;
RUN;
```

gradHS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	91	12.40	91	12.40
1	643	87.60	734	100.00

Binomial Proportion lessHS = 1	
Proportion	**0.8760**
ASE	**0.0122**
95% Lower Conf Limit	0.8522
95% Upper Conf Limit	0.8999
Exact Conf Limits	
95% Lower Conf Limit	0.8500
95% Upper Conf Limit	0.8990

Test of H0: Proportion = 0.9	
ASE under H0	0.0111
Z	**-2.1654**
One-sided Pr > Z	0.0152
Two-sided Pr > |Z|	0.0304

SE for $\boldsymbol{z}$-test uses $\boldsymbol{\sigma}$ instead of $\boldsymbol{s}$

- $\bar{y} = 0.876$, $\mathbf{SE} = 0.0122$ (using $\boldsymbol{s}$)
  - $\mathbf{95\%}\ \boldsymbol{CI} = 0.8522$ to 0.8999
- $\boldsymbol{z} = \frac{0.876-0.90}{0.0111} = -\mathbf{2.165},\ \boldsymbol{p} = \mathbf{.0304}$
  - If $\boldsymbol{\mu} = \mathbf{.90}$, a more extreme sample mean than $\bar{y} = \mathbf{.870}$ (± 2.165 SDs away) would be found ~ **3.04%** of the time
  - 0.876 is smaller than 0.90 (significantly because $p < .05$)

Opposite One-Sample Proportion Test in **STATA**:
Is the proportion of HS graduates different than **.90**?

```
// PRTEST: compare sample proportion to variable=expected at alpha=.05
// STATA always tests proportion of 1 values against H0
prtest gradHS==.90, level(95)
```

```
One-sample test of proportion                   gradHS: Number of obs =      734
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gradHS |   .8760218   .0121642                      .8521805    .8998631
------------------------------------------------------------------------------
    p = proportion(gradHS)                                        z =  -2.1654
Ho: p = 0.9

     Ha: p < 0.9                 Ha: p != 0.9                 Ha: p > 0.9
 Pr(Z < z) = 0.0152         Pr(|Z| > |z|) = 0.0304          Pr(Z > z) = 0.9848
```

- $\bar{\boldsymbol{y}} = 0.876, \mathbf{SE} = 0.0122$ (using $\boldsymbol{s}$); $\mathbf{95\%}\ \boldsymbol{CI} = 0.8522$ to 0.8999
- $\boldsymbol{z} = \frac{0.876-0.90}{0.0111} = \mathbf{-2.165},\ \boldsymbol{p} = \mathbf{.0304}$
  - If $\mu = \mathbf{0.90}$, a more extreme sample mean than $\bar{\boldsymbol{y}} = \mathbf{0.876}$ (± 2.165 SDs away) would be found ~ **3.04%** of the time
  - 0.876 is smaller than 0.90 (significantly because $p < .05$)

Inference via Sampling Distributions

- Two families of options for estimating the **inconsistency of our sample mean** (so far; extensions to sample variance next):

  - Rely on "**asymptotic**" **sampling distributions—what we just did**
    - Asymptotic = what should happen if we had an infinite sample
    - Means using **of-the-shelf formulas** to estimate standard errors
    - A majority of quantitative methods rely on this approach

  - Try to **approximate the sampling distribution** of the statistic through "**resampling**" of the values in the current data
    - Useful when you have a small samples and/or don't have a known sampling distribution that you can rely on for your statistic of interest
    - Basis of techniques like "bootstrapping", "jack-knifing", "permutation tests", and Markov Chain Monte Carlo (MCMC) estimation
    - We won't have time to cover this side (but see ch. 18 of Mitchell 2015)

Wrapping Up

- Salient characteristics of variables → **univariate statistics**:
  - Central tendency (middle of distribution)
    - For categorical variables , is covered by percentage per category
    - For quantitative variables, is covered by mean (and/or median and mode)
  - Dispersion (width of distribution)
    - Dispersion → SD = average deviation from mean (SD^2= variance); also by IQR
    - Skewness (asymmetry) is good to note to guide reporting or analysis

- To **make inferences** from a sample to the population, we need to know how consistent the estimates of the mean and variance would be across samples → this is the **standard error (SE)** of the estimate
  - SE for mean gets smaller (more precise) with more N and less variance
  - We can use t distribution to map obtained sample mean onto expected population mean or create confidence intervals for sampling variation

- See videos for how to use SAS and STATA to get example results!

Bonus Material: Estimating SE of Sample Variance s^2

- We've just seen that the sample mean $\bar{y}$ of sample s will become more normally distributed across samples: $\bar{y}_s \sim \mathrm{N}(\mu, \frac{\sigma}{\sqrt{N}})$ [where (mean, SE)] with increasing N (but will be t-distributed in smaller samples)

- However, the same may not be true of the sample variance s^2
  - Why? The normal and t distributions are continuous and extend from $\pm\infty$ … but what is the smallest number a variance can be?
  - Here's a hint: this picture is an example of what is called a "degenerate distribution"…
  - This is an example distribution for a constant… guess what the variance is for this distribution? 0!

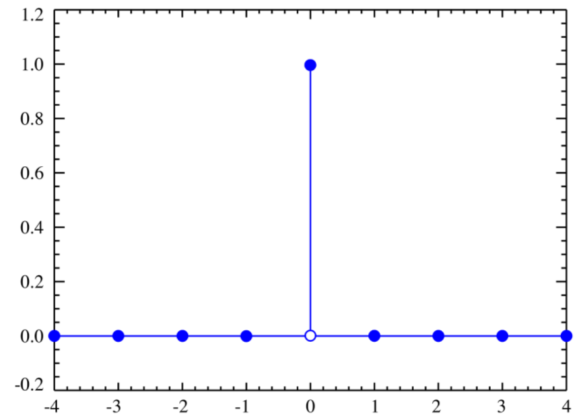


Image borrowed from: https://en.wikipedia.org/wiki/List_of_probability_distributions

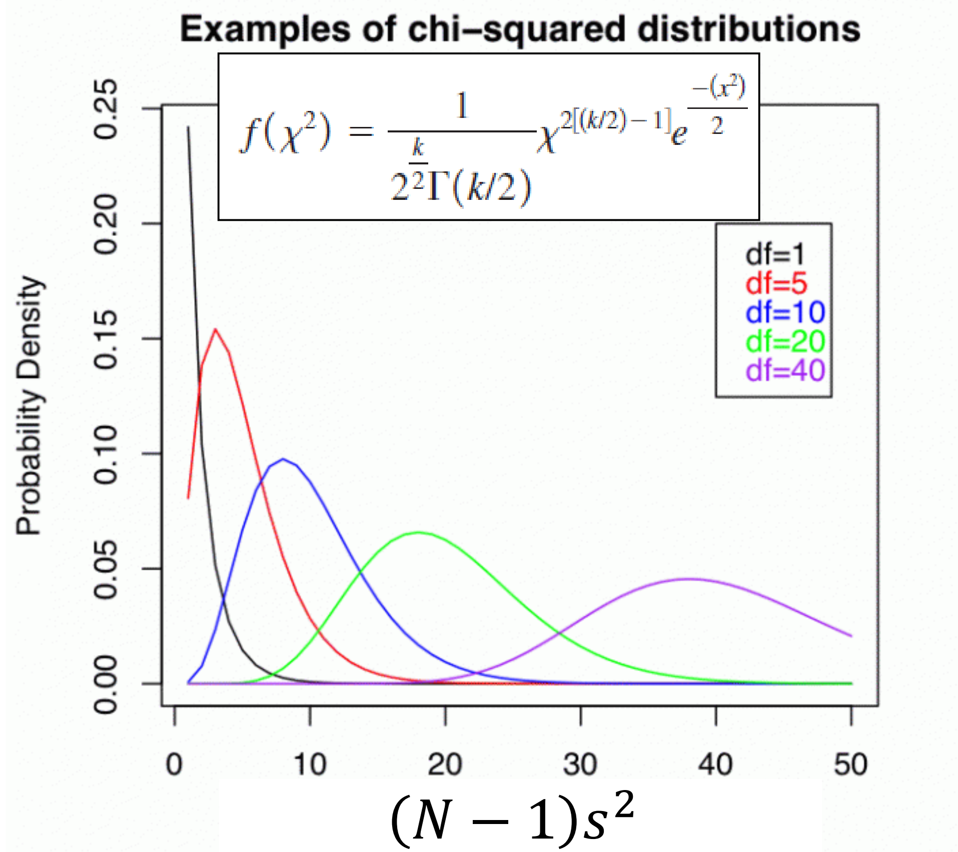
Sampling Distribution for Sample Variance s^2

- Needs to take into account that s^2 must stay above 0
- Which also implies that bigger values of s^2 lend themselves to more variability in what s^2 could be across samples
- If the variable that s^2 refers to is normally distributed, it turns out the χ^2 ("chi-square") distribution works well for this:

$$\frac{(N-1)s^2}{\sigma^2} \sim \chi^2(k = N-1)$$

- The χ^2 distribution has one parameter, k, known as numerator degrees of freedom (more on this soon)

χ^2 Mean $= k$,
χ^2 Variance $= 2k$

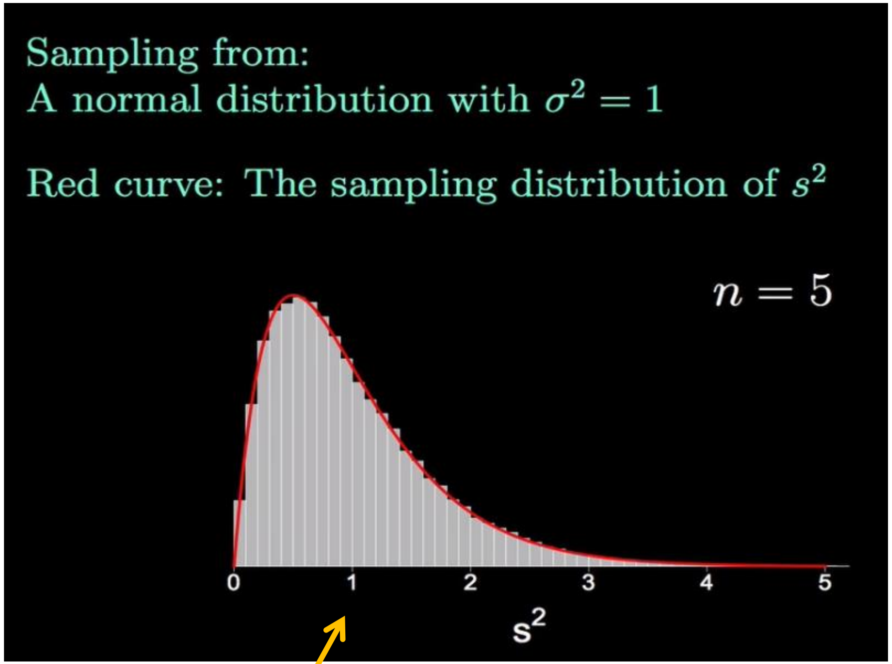


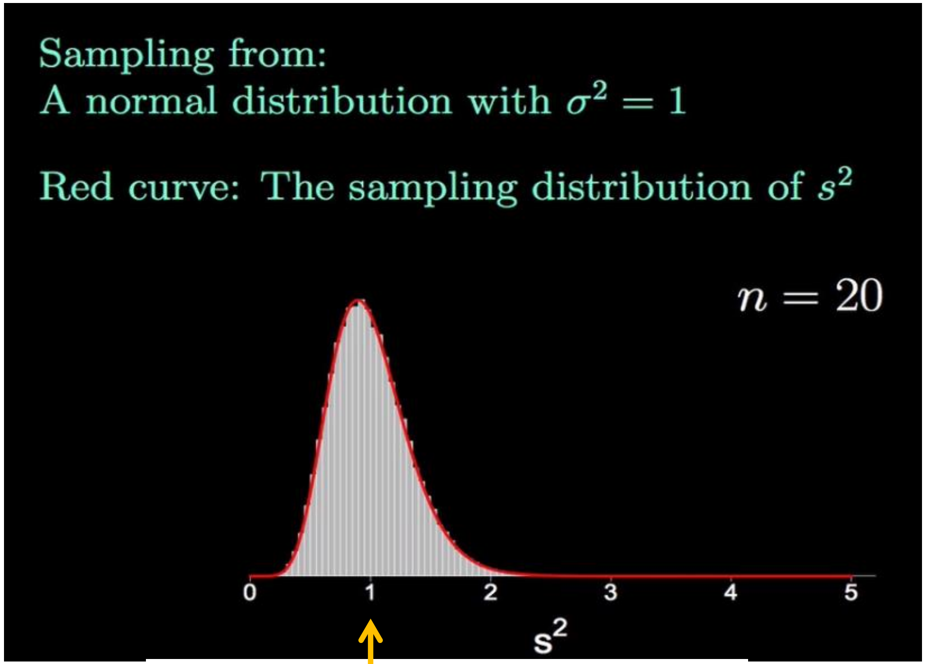
$$\frac{(N-1)s^2}{\sigma^2}$$

Image borrowed from: https://newonlinecourses.science.psu.edu/stat504/node/23/

Sampling Distribution for Sample Variance s^2

- $\frac{(N-1)s^2}{\sigma^2} \sim \chi^2(k = N-1)$ → really non-intuitive way to think about it!
- Let's plot s^2 directly on the x-axis instead:





In smaller samples, the variance is more likely to be underestimated, so the lower boundary at 0 causes skewness

In larger samples, the sampling distribution of the variance is more likely to be symmetric

Images borrowed from: https://www.youtube.com/watch?v=V4Rm4UQHij0/

SE of Sample Variance s^2 (and beyond)

- The χ^2 distribution doesn't hold as closely for variances of other types of variables, but the SE of the variance is not typically of concern in reports of data analysis
- In practice, here's what happens:
  - Statistical software will provide by default the SE for the mean (and for the fixed effects of any model predictor, stay tuned)
  - Software will usually only provide the SE for the variance when using likelihood estimation instead of least squares (as in my other classes)
- Btw, I'm sure there is a way to derive or calculate SEs for other sample statistics (median, mode, skewness, kurtosis), but I've never once needed to do so…
  - Resampling approaches (e.g., bootstrapping) are likely your best bet