Confidence Intervals and Hypothesis Testing

Confidence Intervals

"Information derived from a sample of observations can only be an approximation of the entire population. To make a definitive statement about an entire population , every member of that population would need to be measured."

Myatt, Glenn J., and Wayne P. Johnson. *Making Sense of Data I : A Practical Guide to Exploratory Data Analysis and Data Mining*, John Wiley & Sons, Incorporated, 2014. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/allegheny-ebooks/detail.action?docID=1729064.

- Not everything can be measured
- We need to make estimates based on existing data
- This leads to ability to say something is "statistically significant"

Sample vs Observation

"Confidence intervals are a measure of our uncertainty about the statistics we calculate from a **single sample** of observations."

- Nb, a sample is a set of observations
- If every possible observation were recorded, we would no longer have a sample, we would have an entire population

- e.g. one sample of 10 humans → 10 weights are observed → single sample mean

Sample vs Population

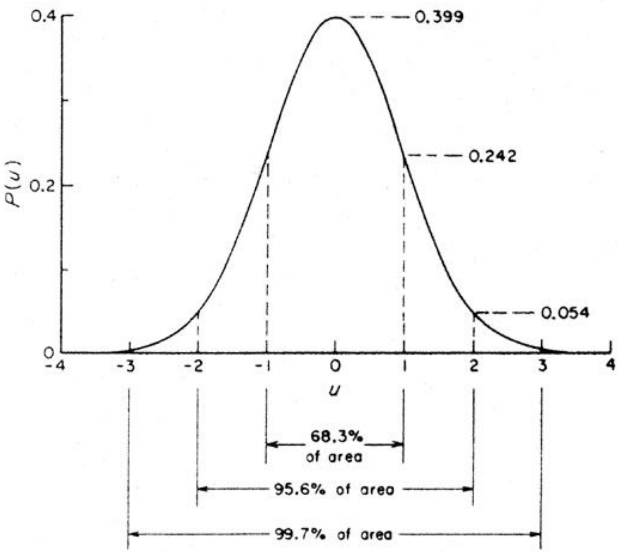
- e.g. one sample of 10 humans → 10 weights are observed → single sample mean
- 10 humans is not the entire population

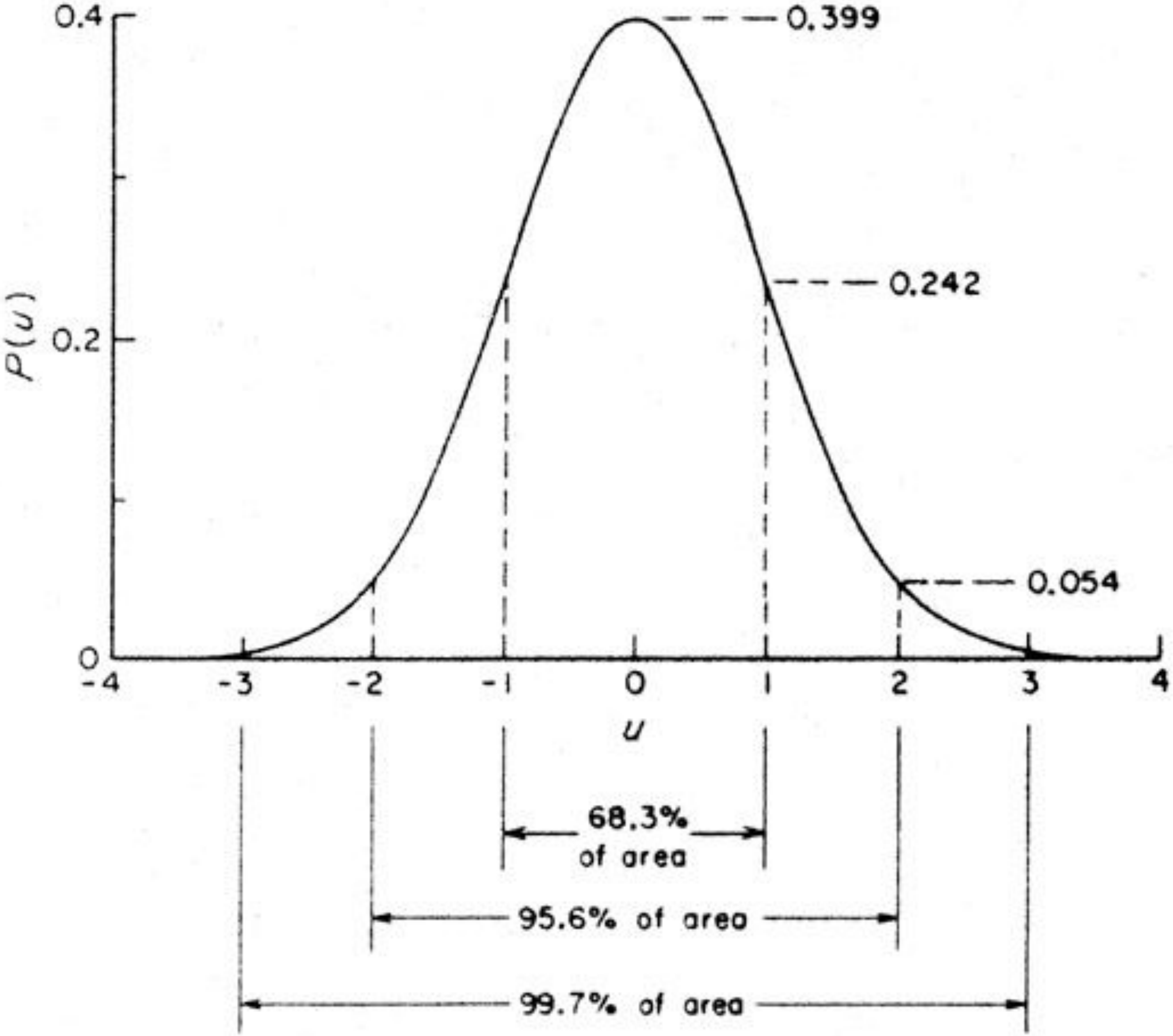
- A different sample would have a slightly different mean
- yet a different sample would have a slightly different mean
- The distribution of means is called the **sampling distribution** of the mean
- The mean of the sampling distribution is called the **expected value** of the mean
- The standard dev of sampling distribution is the **standard error** of the mean

Thanks to the central limit theorem

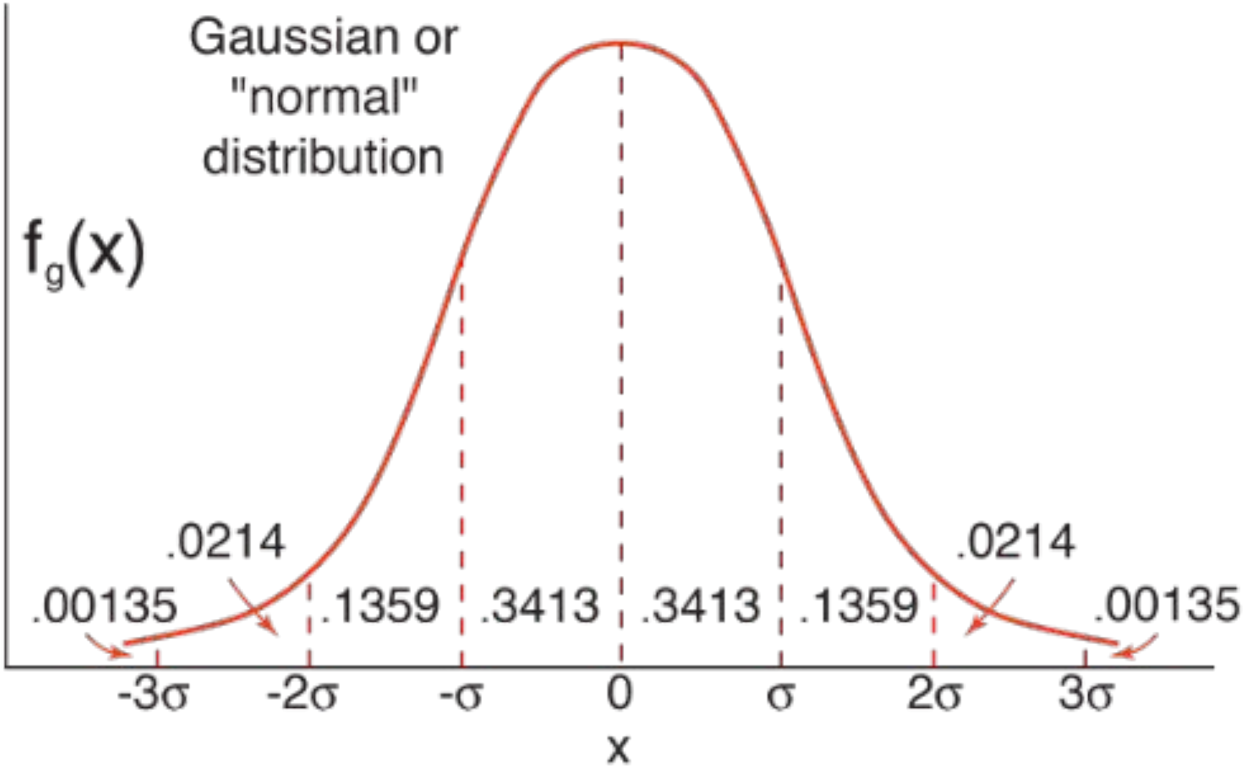
- sampling distributions are gaussian! (normal, bell)
- the properties of gaussian distributions are very nice
- https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

- The normalized area under the curve is 1
- 95% of the observations fall within ±2 sigma!





0.4
0.2
0
P(u)
0.399
0.242
0.054
-4
-3
-2
-1
0
1
2
3
4
u
68.3% of area
95.6% of area
99.7% of area


Gaussian or
"normal"
distribution
fg(x)
.0214
.00135
.1359
.3413
.3413
.1359
.0214
.00135
-3σ
-2σ
-σ
0
σ
2σ
3σ
x

Connecting Observations to Populations

Formula for Standard Error of the Mean:

- SEM = SD / $\sqrt{n}$
  - SD is standard deviation of a sample
  - n is number of observations in sample

- For fixed SD
  - if n is small, SEM is ?
  - if n is large, SEM is ?

Connecting Observations to Populations

Formula for Standard Error of the Mean:

- $SEM = SD / \sqrt{n}$
  - SD is standard deviation of a sample
  - n is number of observations in sample

- For fixed SD
  - if n is small, SEM is large
  - if n is large, SEM is small

Connecting Observations to Populations

If we only measure the weight of 30 people, but we want to report a number range for which we are 95% sure to include the REAL population mean for weight…

- confidence interval → 95 %
- alpha = 0.05
- alpha / 2 = 0.025
- actual range given by:
  - s is observed standard dev
  - n is number of observations
  - Critical Z value → 1.96

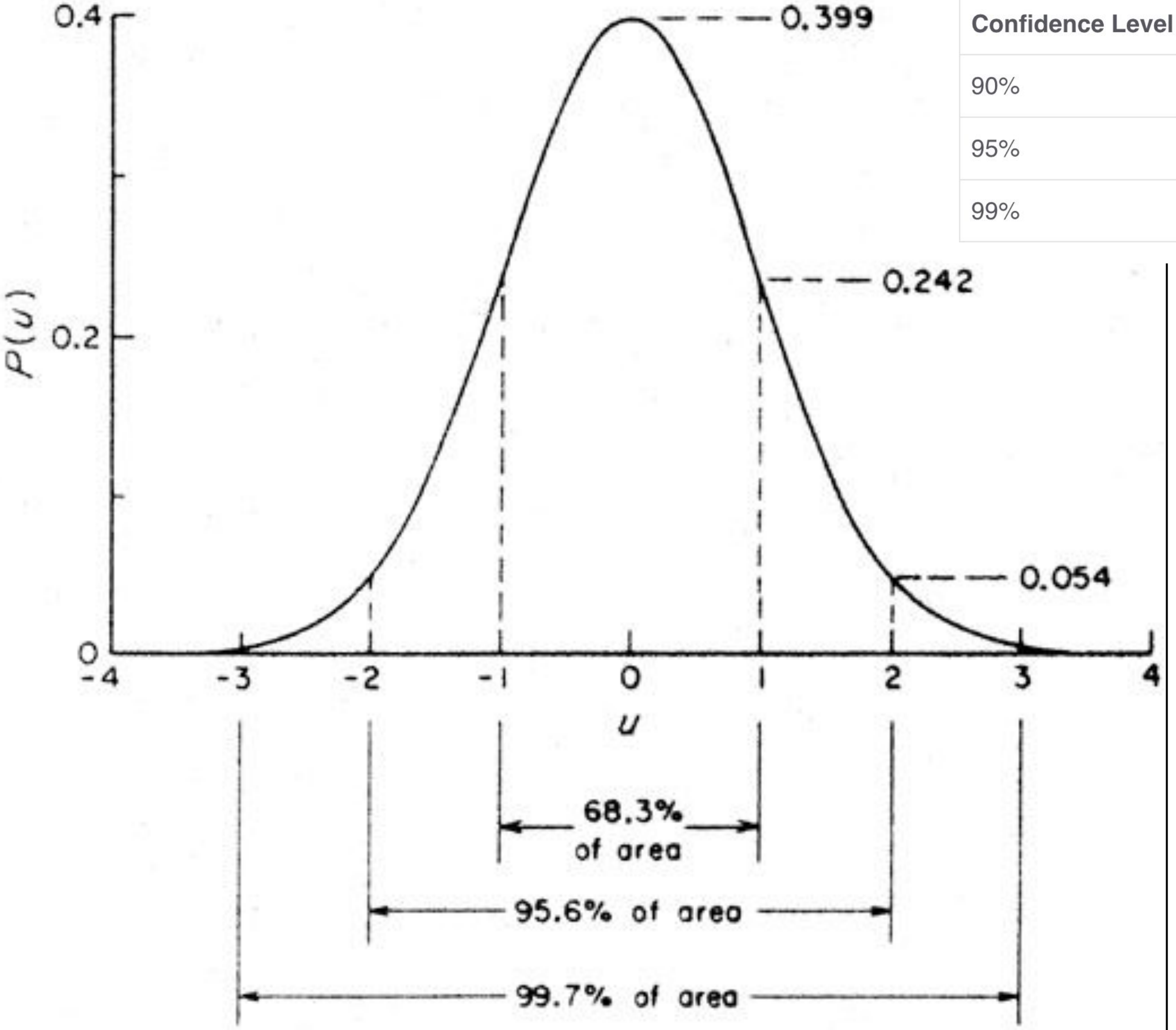
$$\textit{confidence interval} = 100 \times (1 - \alpha)$$

$$\bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Nb. Critical Z is related to the concept of z-scored data

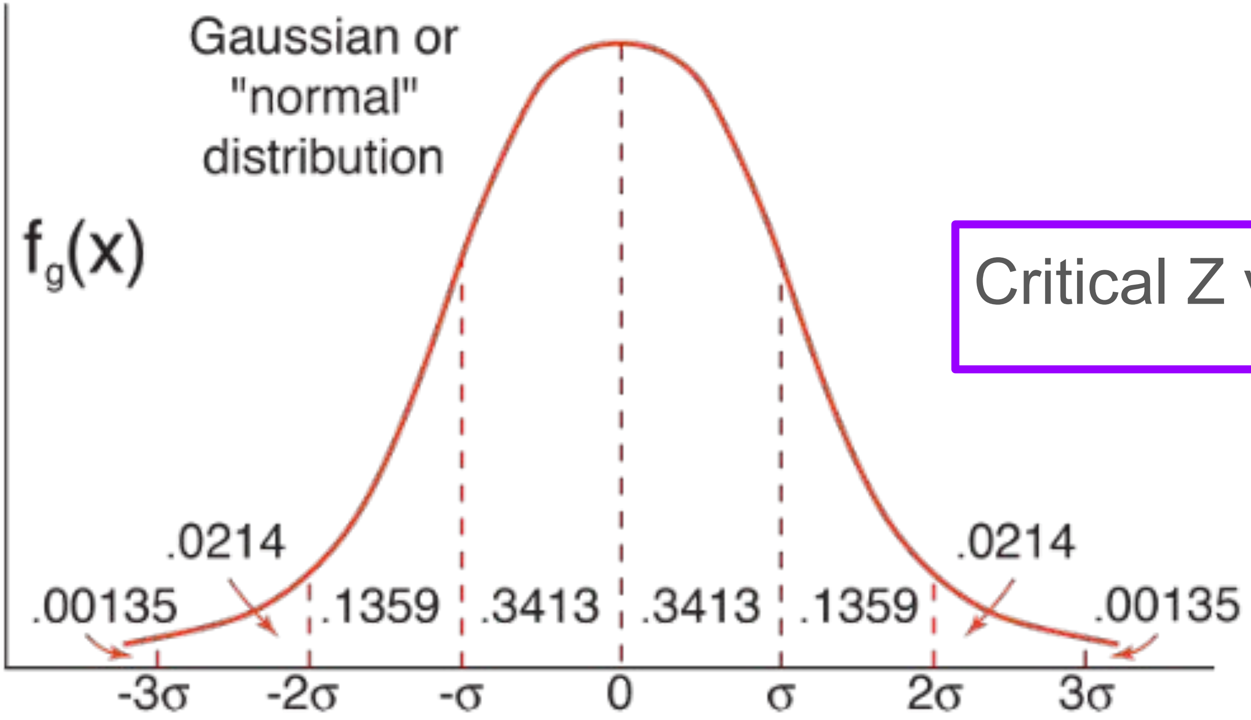
$$z = \frac{x_i - \bar{x}}{s}$$

- measures data relative to mean and SD



Confidence Level	Two Sided CV
90%	1.64
95%	1.96
99%	2.58

Critical Z values



Critical Z values

Confidence Level	Two Sided CV	One Sided CV
90%	1.64	1.28
95%	1.96	1.65
99%	2.58	2.33

Fun way to find critical Z

- use numpy to make data that is drawn from a proper gaussian distribution
- find the value of the 97.5 percentile!
- ^^^ is for alpha level of 0.05: 1 - 0.05/2 = 97.5

```
#%% Calculate Critcal Z

alpha = 0.05
# Make gaussian data
mu, sigma = 0, 1 # sample mean and SEM
normdata = np.random.normal(mu, sigma, 10000000)
print(np.quantile(normdata, 1 - alpha/2))
```

Hypothesis Testing

Saying with some degree of certainty that a baseline hypothesis is true or false

- example, the weight of 100 shampoo bottles are observed…
  - mean is 199.94g
  - SD is 0.613
    - recall SD is square root of variance;
- baseline, or null hypothesis is that the **sampling average weight** is 200g
- problem: determine if the observed mean would reasonably be observed if the true mean were 200g
  - let the confidence level be set to 95%
  - Critical Z for 95% percent confidence is ±1.96
  - Use the observed mean to compute the **standardized test statistic**
  - If T is within the bounds we choose, the baseline null hypothesis stands!

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Hypothesis Testing

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Saying with some degree of certainty that a baseline hypothesis is true or false

- example, the weight of 100 shampoo bottles are observed…
  - mean is 199.94g
  - SD is 0.613
    - recall SD is square root of variance;
- baseline, or null hypothesis is that the **sampling average weight** is 200g
- problem: determine if the observed mean would reasonably be observed if the true mean were 200g
  - let the confidence level be set to 95%
  - Critical Z for 95% percent confidence is ±1.96
  - Use the observed mean to compute the **standardized test statistic**
  - If T is within the bounds we choose, the baseline null hypothesis stands!

T = (199.94 - 200) / (0.613 / 10)

T = -0.979

CI = ± 1.96

T is within the CI, therefore the baseline hypothesis is not rejected!

The observed data do not disprove the hypothesis that 200g is the true weight of a shampoo bottle.

For observations > 30 !

- compared to 0

$$T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

$$T = \frac{\bar{x}_1 - \mu_1}{\sqrt{\frac{s_1^2}{n_1}}}$$

- compared to independent sample

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$