Exam 2 Review

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Copyright D.B. Rowe

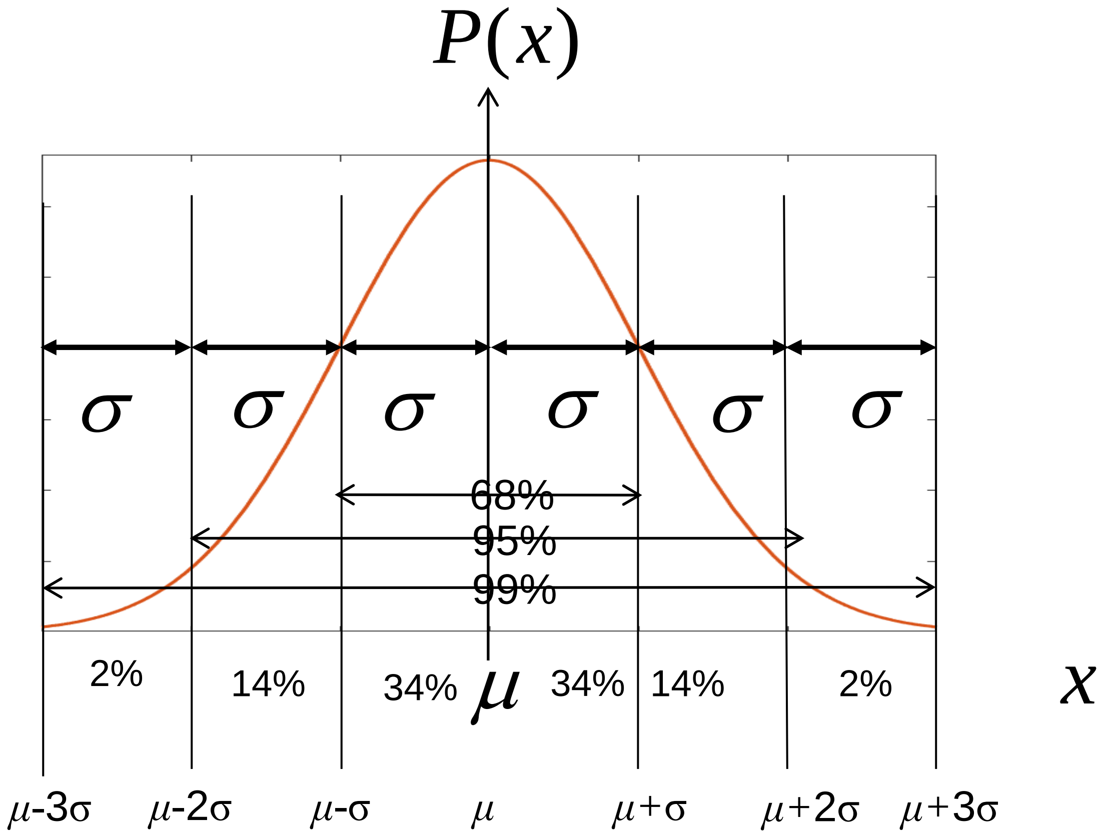
5.6 Probability Models – Normal Distribution

The normal distribution is often used for continuous outcomes.

You may know it as the bell curve or Gaussian distribution.

Its functional form is

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Total Area Under Curve = 1

Symmetric about the mean.

mean = median = mode.

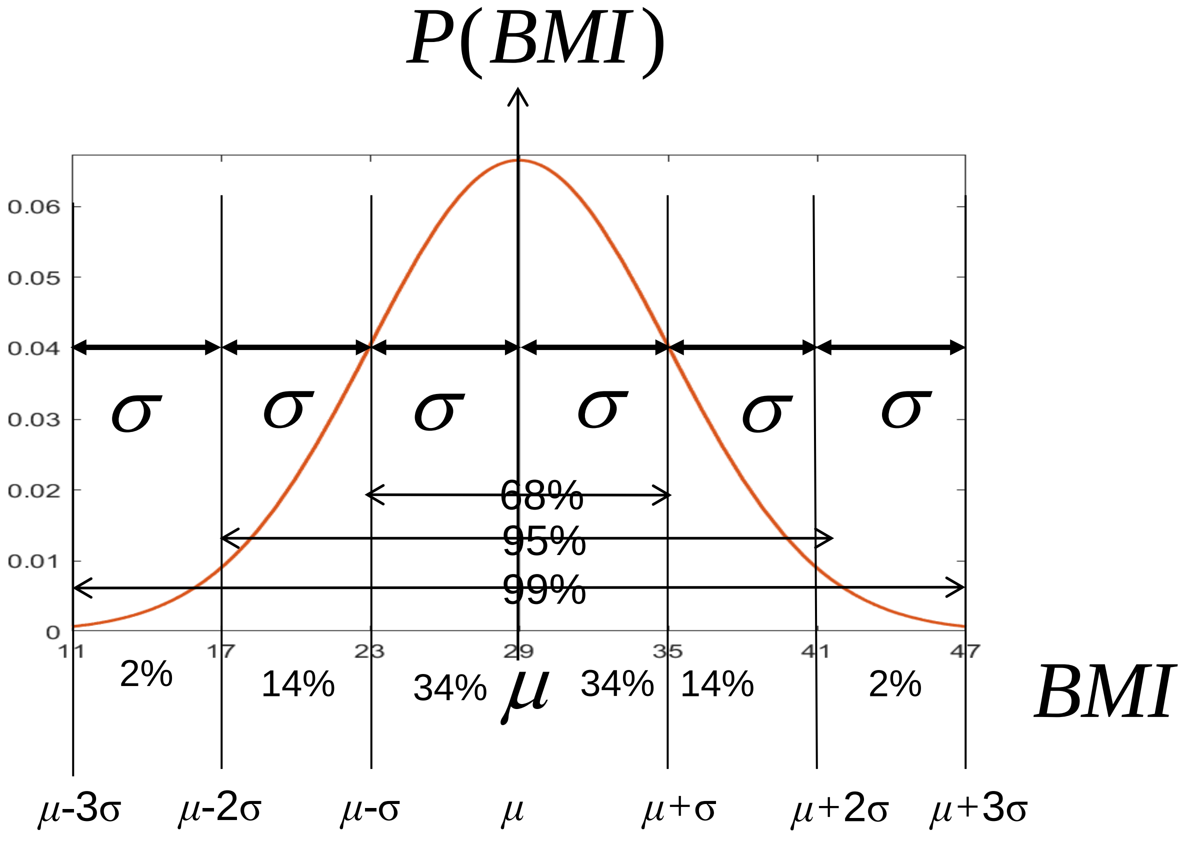
mean μ & variance σ^2

5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

Its functional form is

$$P(x)=\frac{1}{6\sqrt{2\pi}}e^{-\frac{(x-29)^2}{2(6)^2}}$$



Total Area Under Curve = 1

Symmetric about the mean.

mean = median = mode.

mean μ & variance σ^2

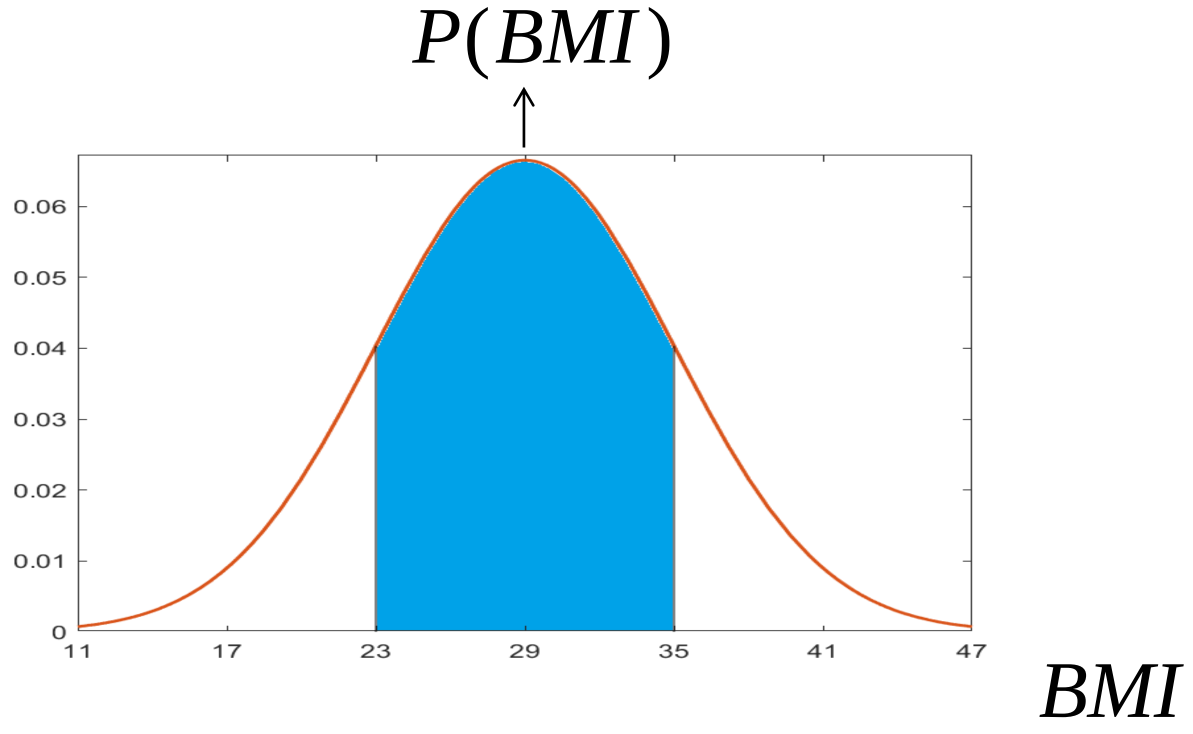
5.6 Probability Models – Normal Distribution

The mean BMI for males aged 60 is μ=29 kg/m^2 with standard deviation σ=6 kg/m^2 (with a normal distribution).

Normally in math we do something called an integral. $x=BMI$

$$A = P(23 < x < 35) = \int_{23}^{35} \frac{1}{6\sqrt{2\pi}} e^{-\frac{(x-29)^2}{2(6)^2}} dx$$

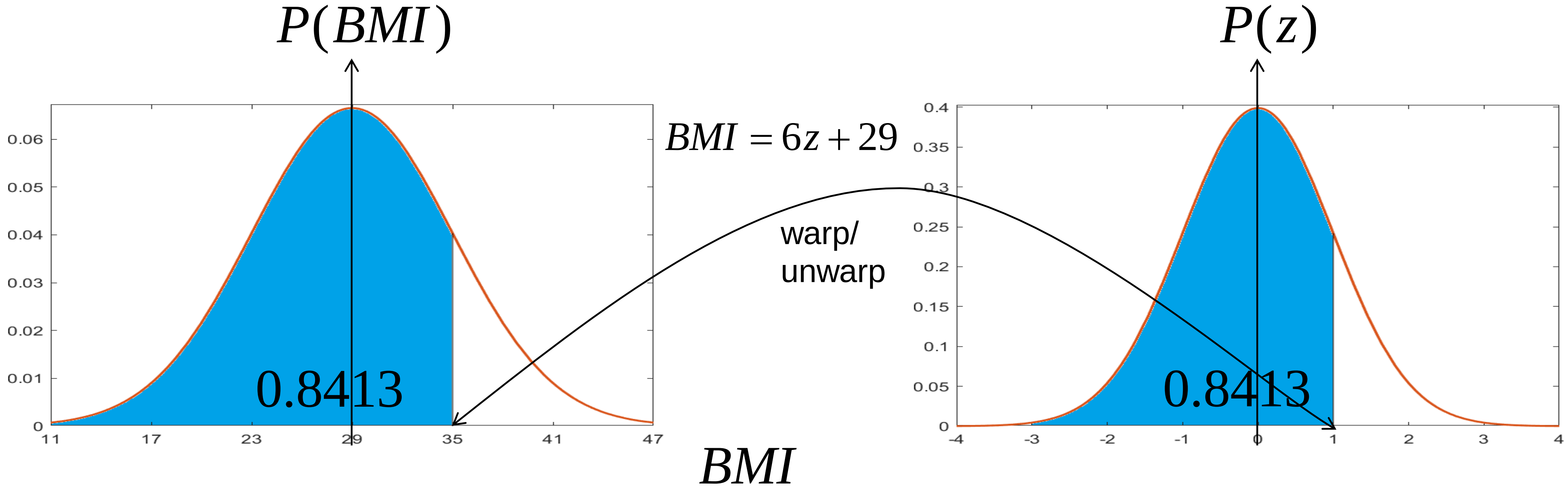
But we are not doing Calculus and even if we know Calculus, we can't integrate $P(x)$!



Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

We need to convert from the BMI (x) axis to a new "z" axis, $z = \dfrac{x-\mu}{\sigma}$.



$P(BMI<35)$ $P(z<1)$

z_i	.00	.01
0.0	0.5000	0.5040
0.1	0.5398	0.5438
0.2	0.5793	0.5832
0.3	0.6179	0.6217
0.4	0.6554	0.6591
0.5	0.6915	0.6950
0.6	0.7257	0.7291
0.7	0.7580	0.7611
0.8	0.7881	0.7910
0.9	0.8159	0.8186
1.0	0.8413	0.8438

A

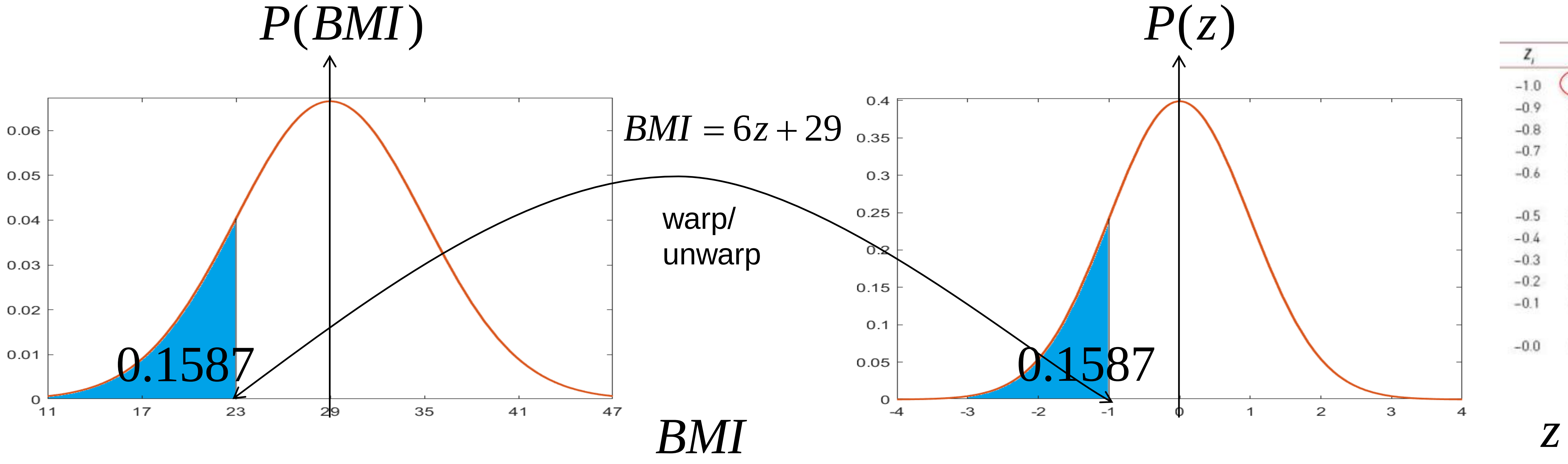
z

Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

5.6 Probability Models – Normal Distribution

We need to convert from the BMI (x) axis to a new “z” axis, $z = \dfrac{x-\mu}{\sigma}$.



A

z_i	.00	.01
−1.0	0.1587	0.1562
−0.9	0.1841	0.1814
−0.8	0.2119	0.2090
−0.7	0.2420	0.2389
−0.6	0.2743	0.2709
−0.5	0.3085	0.3050
−0.4	0.3446	0.3409
−0.3	0.3821	0.3783
−0.2	0.4207	0.4168
−0.1	0.4602	0.4562
−0.0	0.5000	0.4960

$P(BMI<23)$ $P(z<-1)$

Area under curve on z axis same as area under curve on x axis.

Total Area Under Curve = 1

5.6 Probability Models – Sampling Distributions

One major thing is to take a random sample $x_1,\ldots,x_n$, and average, $\bar{X}$.

The **Sampling Distribution** says, if $x_1,\ldots,x_n$, is from a population with mean μ and standard deviation σ, then $\bar{X}$ has $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma / \sqrt{n}$.
So by averaging, we've reduced our standard deviation!

Above the Sampling Distribution is the **Central Limit Theorem (CLT)**.
The **CLT** says, that if n is large, i.e. $n>30$, then $\bar{X}$ has an approximately normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma / \sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

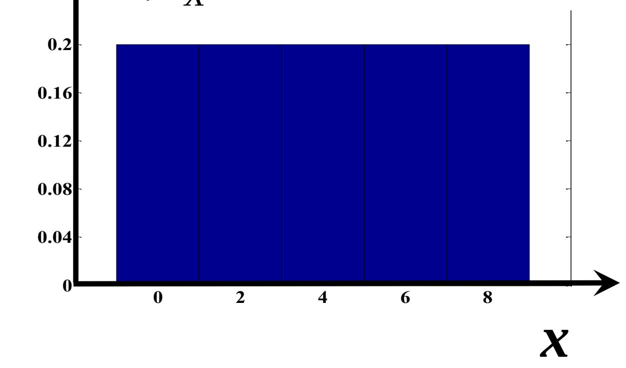
This is **HUGE**, meaning we can use our old friend the normal distribution.

5.6 Probability Models – Sampling Distributions

Example: N=5 balls in bucket, selecting increasing n with replacement.

$\bar{X}$ from n=1 distribution

$\mu_{\bar{X}} = 4, \sigma = \sqrt{8}$



$\bar{X}$ from n=2 distribution

$\mu_{\bar{X}} = 4, \sigma_{\bar{X}} = \sqrt{8}/\sqrt{2}$

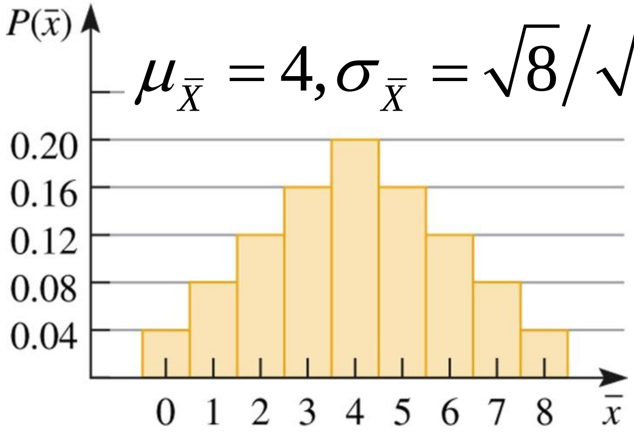
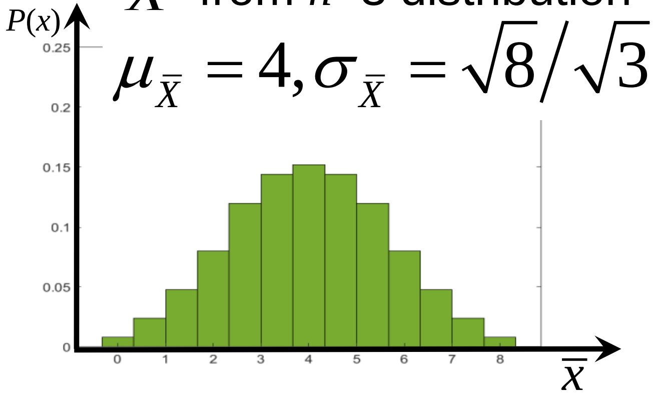


Figure from Johnson & Kuby, 2012.

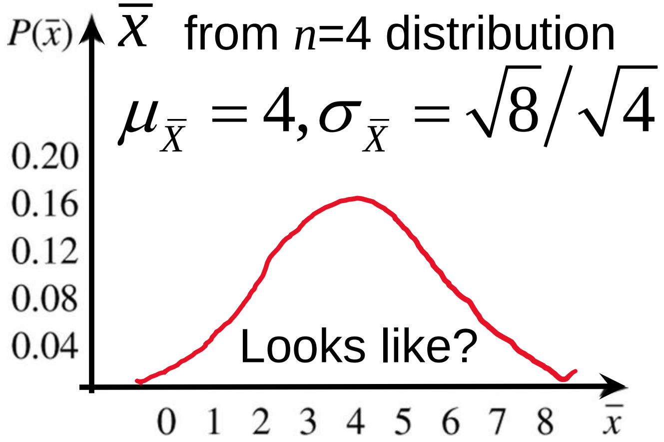
$\bar{X}$ from n=3 distribution

$\mu_{\bar{X}} = 4, \sigma_{\bar{X}} = \sqrt{8}/\sqrt{3}$



$\bar{X}$ from n=4 distribution

$\mu_{\bar{X}} = 4, \sigma_{\bar{X}} = \sqrt{8}/\sqrt{4}$



$\xrightarrow{n \text{ large?}}$

$\mu_{\bar{X}} = 4$

$\sigma_{\bar{X}} = \sqrt{8}/\sqrt{n}$

Looks like?

The **Sampling Distribution** says, if we take a random sample $x_1,\ldots,x_n$, from a population with mean μ and standard deviation σ and average the observations , then the average has a mean μ and standard deviation $\sigma/\sqrt{n}$. So by averaging, we've reduced our standard deviation!

The **Central Limit Theorem** (**CLT**) says, that if n is large, i.e. $n>30$, then the average has an approximately normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ no matter what original distribution the data $x_1,\ldots,x_n$ came from.

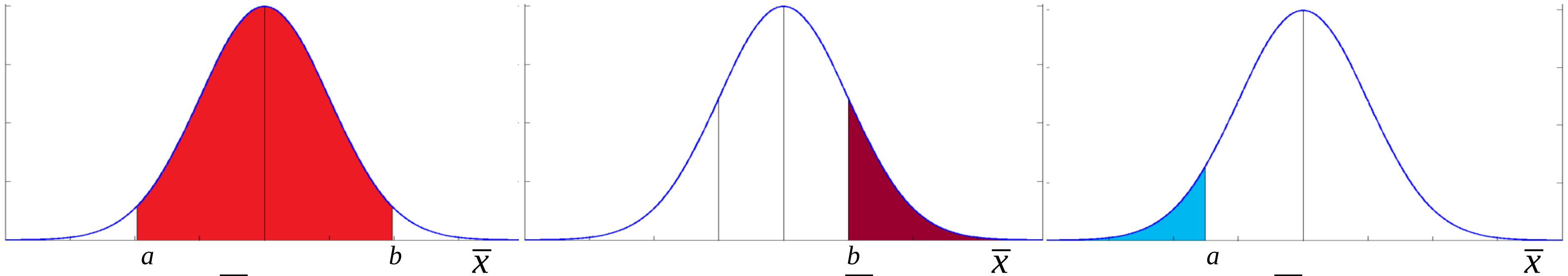
5.6 Probability Models – Normal Distribution

Now that we know that $\bar{X}$ has a normal distribution when n is large, we can find probabilities (areas) for finding a random mean by converting to a z and using the tables.

$$x_1,\ldots,x_n$$

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



$P(a < \bar{X} < b)$ same area $P(c < z < d)$

$P(b < \bar{X})$ same area $P(d < z)$

$P(\bar{X} < a)$ same area $P(z < c)$

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad c = \frac{a - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad d = \frac{b - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

Example:
What is probability that sample mean $\bar{X}$ from a random sample of n=16 heights is greater than 70" when μ =67 and σ =4?

6.1 Introduction to Estimation

A **Point Estimate** for a population parameter is a single-valued estimate of that parameter. i.e. $\bar{X}$ for μ or s^2 for σ^2.

A **Confidence Interval** (CI) estimate is a range of values for a population parameter with a confidence attached (i.e., 95%).

A CI starts with the Point Estimate and builds in what is called the **Margin of Error**. The margin of error incorporates probabilities.

$$\underbrace{\bar{X}}_{\text{PE}} \pm \underbrace{\textit{that depends on a probability}}_{\text{ME}}$$

6.2 Confidence Intervals for One Sample, Continuous Outcome

Example: Find the value of $t_{0.025,10}$.

$$\underbrace{0.025}_{\alpha/2},\ \underbrace{10}_{df=n\text{-}1}$$

The (critical) value of t that has an area of 0.025 larger than it when we have 10 degrees of freedom is 2.228.

This is the value we use for a 95% CI when α=0.05 and n=11.

Book says $n \geq 30$ use bottom z value.

Confidence Level	80%	90%	95%	98%	99%
Two-Sided Test α	.20	.10	.05	.02	.01
One-Sided Test α	.10	.05	.025	.01	.005
$df=n\text{-}1$					
1	3.078	6.314	12.71	31.82	63.66
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
z ∞	1.282	1.645	1.960	2.326	2.576

6.2 Confidence Intervals for One Sample, Continuous Outcome

$100(1\text{-}\alpha)\%$

Example: Suppose we wish to compute a 95% CI for true systolic BP.
A random sample of n=10 is take with sample mean $\bar{X}=121.2$ mm Hg
and sample standard deviation $s=11.1$ mm Hg.
The equation (when σ unknown) is $\bar{X} \pm t_{\frac{\alpha}{2},df} \frac{s}{\sqrt{n}}$, $df=n$-1.
We find the critical t value in the table.
Down to row df=9 and over to column
CI=95% (or two side-test α=0.05,
or one side-test α=0.025).

$$121.2 \pm (2.262)\frac{11.1}{\sqrt{10}} \rightarrow 113.3 \ to \ 129.1$$
mm Hg

n=10, df=9 and α=0.05

Confidence Level	80%	90%	95%	98%	99%
Two-Sided Test α	.20	.10	.05	.02	.01
One-Sided Test α	.10	.05	.025	.01	.005
df					
1	3.078	6.314	12.71	31.82	63.66
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

6.4 Confidence Intervals for Two Independent Samples, Continuous Outcome

$$S_P^2 = \frac{(n_1-1)s_1^{\,2}+(n_2-1)s_2^{\,2}}{n_1+n_2-2}$$

Example: A sample of n=10 males and females had systolic blood pressure measured. The data are: males: n_1=6, $\bar{X}_1$=117.5 mm Hg s_1=9.7 mm Hg and females n_2=4, $\bar{X}_2$=126.8 mm Hg, s_2=12 mm Hg.

Generate a 95% CI for μ_1-μ_2.

$$\bar{X}_1-\bar{X}_2 \pm t_{\frac{\alpha}{2},df} S_P\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad , \quad df = n_1+n_2-2 \quad , \quad S_P = \sqrt{\frac{(n_1-1)(s_1)^2+(n_2-1)(s_2)^2}{n_1+n_2-2}}$$

from table

$$df = 6+4-2 = 8 \qquad S_P = \sqrt{\frac{(6-1)(9.7)^2+(4-1)(12.1)^2}{6+4-2}} = 10.6 \text{ mm Hg}$$

$$(117.5-126.8)\pm(2.306)(10.6)\sqrt{\frac{1}{6}+\frac{1}{4}} \longrightarrow -9.3\pm 15.78 \text{ mm Hg} \longrightarrow -25.08 \;\; to \;\; 6.48 \text{ mm Hg}$$

from table

6.4 Confidence Intervals for Two Independent Samples, Continuous Outcome

Example: Difference in Systolic blood pressure between two visits.

Subject Identification Number	Examination 6	Examination 7	Difference	Difference – $\bar{X}_d$	(Difference – $\bar{X}_d$)^2
1	168	141	–27	–21.7	470.89
2	111	119	8	13.3	176.89
3	139	122	–17	–11.7	136.89
4	127	127	0	5.3	28.09
5	155	125	–30	–24.7	610.09
6	115	123	8	13.3	176.89
7	125	113	–12	–6.7	44.89
8	123	106	–17	–11.7	136.89
9	130	131	1	6.3	39.69
10	137	142	5	10.3	106.09
11	130	131	1	6.3	39.69
12	129	135	6	11.3	127.69
13	112	119	7	12.3	151.29
14	141	130	–11	–5.7	32.49
15	122	121	–1	4.3	18.49
			-79.6	0.5	2296.95

$$\bar{X}_d \pm t_{\frac{\alpha}{2},df} \frac{s_d}{\sqrt{n}}$$

from table

Compute a 95% CI.

$$\bar{X}_d = \frac{-79.0}{15} - 5.3 \text{ mm Hg}$$

$$s_d = \sqrt{\frac{2296.95}{15-1}} = \sqrt{164.07} = 12.8 \text{ mm Hg}$$

$df=n$-1

Confidence Level	80%	90%	95%
Two-Sided Test α	.20	.10	.05
One-Sided Test α	.10	.05	.025
df			
1	3.078	6.314	12.71
2	1.886	2.920	4.303
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131

$$-5.3 \pm (2.145)\frac{12.8}{\sqrt{15}} \longrightarrow -5.3 \pm 7.1 \longrightarrow -12.4 \text{ to } 1.8 \text{ mm Hg}$$

from table

6.6 Confidence Intervals for Two Independent Samples, Dichotomous Outcome

$$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

The CI for the natural log of relative risk, $ln(RR)$ is:

$$ln(RR) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{(n_1 - X_1)/X_1}{n_1} + \frac{(n_2 - X_2)/X_2}{n_2}}$$

CI for relative risk (RR) is:

$$\exp(Lower\ Limit), \exp(Upper\ \ Limit)$$

We go through the same process.

6.6 Confidence Intervals for Two Independent Samples, Dichotomous Outcome

$$OR = \frac{\hat{p}_1 / (1-\hat{p}_1)}{\hat{p}_2 / (1-\hat{p}_2)}$$

CI for the natural log of odds ratio, $ln(OR)$ is:

$$ln(OR) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{X_1} + \frac{1}{n_1 - X_1} + \frac{1}{X_2} + \frac{1}{n_2 - X_2}}$$

CI for odds ratio, OR is:

$$\exp(Lower\ Limit), \exp(Upper\ \ Limit)$$

We go through the same process.

6.7 Summary

Number of Groups, Outcome: Parameter	Confidence Interval, $n<30$	Confidence Interval, $n\geq 30$
One sample, continuous: CI for μ	$\bar{X} \pm t_{\frac{\alpha}{2},df} \frac{s}{\sqrt{n}}$	$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$
One sample, dichotomous: CI for p	(Not taught in this class.)	$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Two independent samples, continuous: CI for μ_1-μ_2	$(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2},df} S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $S_P = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ $df = n_1 + n_2 - 2$	$(\bar{X}_1 - \bar{X}_2) \pm z_{\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $S_P = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
Two matched samples, continuous: CI for μ_d=μ_1-μ_2	$\bar{X}_d \pm t_{\frac{\alpha}{2},df} \frac{s_d}{\sqrt{n}}$	$\bar{X}_d \pm z_{\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$
Two independent samples, dichotomous: CI for RD=(p_1-p_2)	(Not taught in this class.)	$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
CI for $ln(RR)$=$ln(p_1/p_2)$	(Not taught in this class.)	$\ln(RR) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{(n_1-X_1)/X_1}{n_1} + \frac{(n_2-X_2)/X_2}{n_2}}$
CI for RR=p_1/p_2	(Not taught in this class.)	$\exp(Lower\ Limit), \exp(Upper\ \ Limit)$
CI for $ln(OR)$= $ln([p_1/(1-p_1)]/[p_2/(1-p_2)])$	(Not taught in this class.)	$ln(OR) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{X_1} + \frac{1}{n_1-X_1} + \frac{1}{X_2} + \frac{1}{n_2-X_2}}$
CI for OR=$[p_1/(1-p_1)]/[p_2/(1-p_2)]$	(Not taught in this class.)	$\exp(Lower\ Limit), \exp(Upper\ \ Limit)$

Hypothesis Testing

We make decisions every day in our lives.

Should I believe A or should I believe B (not A)?

Two Competing Hypotheses. A and B.

Null Hypothesis ($\mathbf{H_0}$): No difference, no association, or no effect.

Alternative Hypothesis ($\mathbf{H_1}$): Investigators belief.

The Alternative Hypothesis is always set up to be what you want to build up evidence to prove.

7.1 Introduction to Hypothesis Testing

The hypothesis testing process consists of 5 **Steps**.

Step 1: Set up the hypotheses and determine the level of significance.

State the null and the alternative hypotheses.

H_0: Null Hypothesis (no change, no difference)

vs.

H_1: Research Hypothesis (investigators belief, what we want to prove)

Select a level of significance α. $\alpha=0.05$

7.1 Introduction to Hypothesis Testing

The hypothesis testing process consists of 5 Steps.

Step 1: Set up the hypotheses and determine the level of significance.

There are three possible pairs. $\alpha=0.05$

$\text{H}_0\text{: } \mu = \mu_0$ *vs*. $\text{H}_1\text{: } \mu > \mu_0$ (prove greater than, **upper tailed test**)

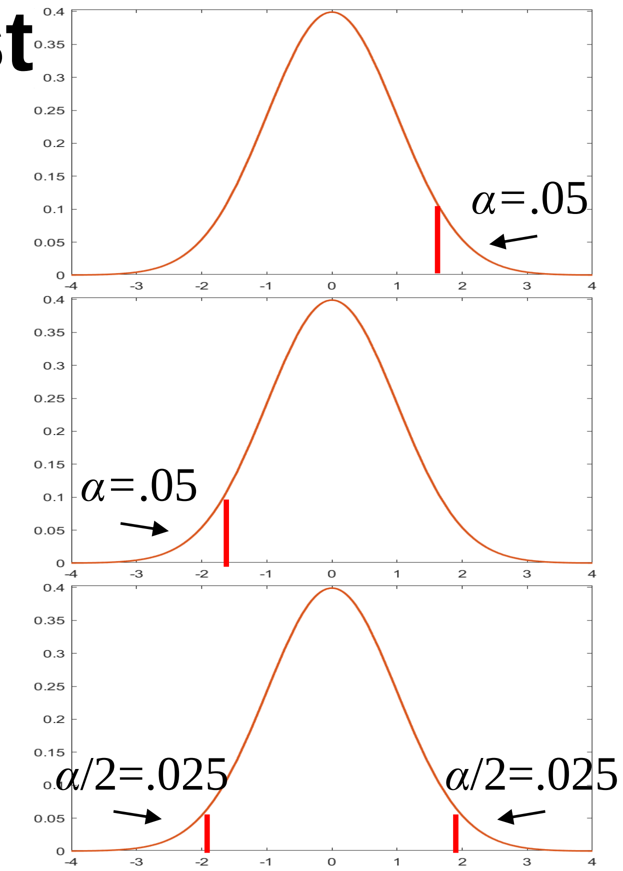
$\quad\le$ reject for "large" $\bar{X}$ or z's

$\text{H}_0\text{: } \mu = \mu_0$ *vs*. $\text{H}_1\text{: } \mu < \mu_0$ (prove less than, **lower tailed test**)

$\quad\ge$ reject for "small" $\bar{X}$ or z's

$\text{H}_0\text{: } \mu = \mu_0$ *vs*. $\text{H}_1\text{: } \mu \neq \mu_0$ (prove not equal to, **two-tailed test**)

reject for "large" or "small" $\bar{X}$ or z's



7.1 Introduction to Hypothesis Testing

The hypothesis testing process consists of 5 Steps.

Step 2: Select the appropriate test statistic.

The test statistic is a single (decision) number.

n large

$$z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

n small

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \qquad df=n\text{-}1$$

Use the test statistic that depends on data and null hypothesis with a critical value z_a (or $t_{a,df}$) that depends on significance level α to make decision.

$a = \alpha$ or $\alpha/2$

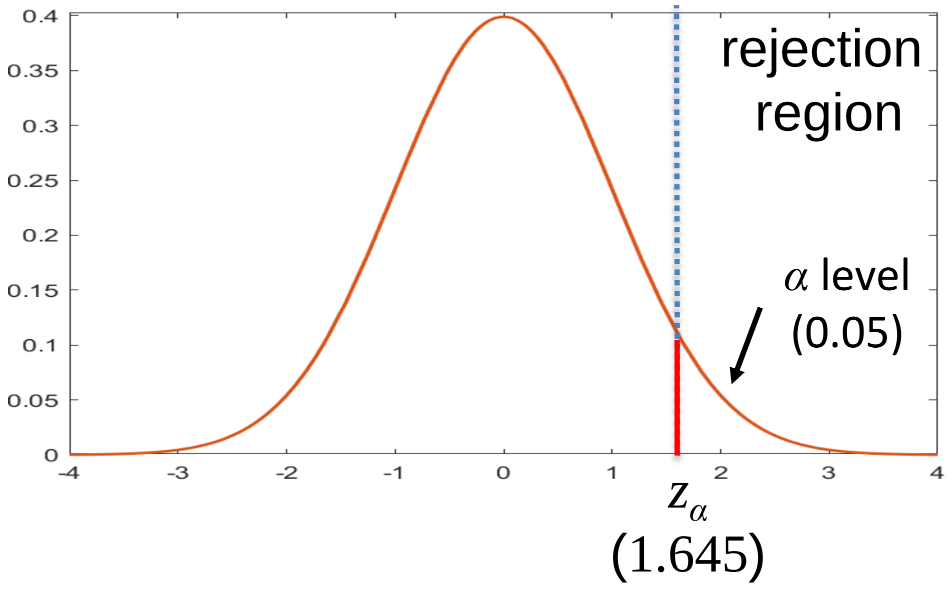
We will test hypotheses on various parameters with various test statistics.

7.1 Introduction to Hypothesis Testing

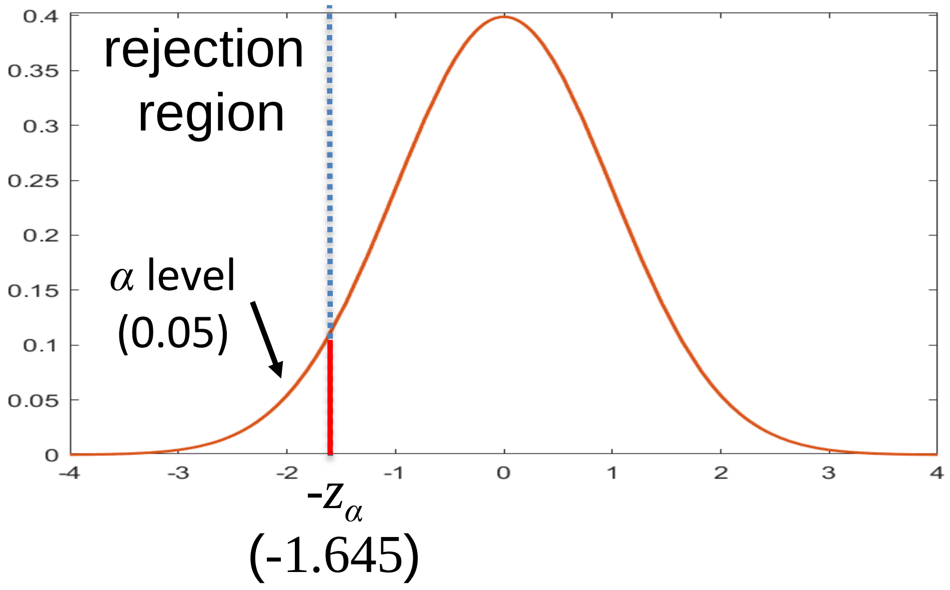
The hypothesis testing process consists of 5 Steps.

Step 3: Set-up the decision rule.

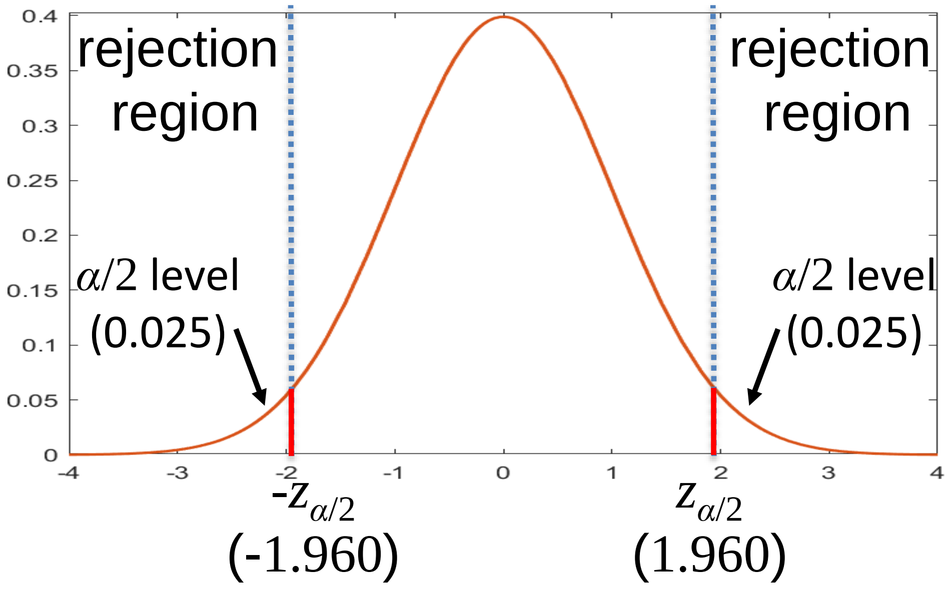
$H_0: \mu = \mu_0$ vs. $H_1: \mu > \mu_0$



$H_0: \mu = \mu_0$ vs. $H_1: \mu < \mu_0$



$H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$



Reject H_0 if $z \geq z_\alpha$

Reject H_0 if $z \leq -z_\alpha$

Reject H_0 $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$

7.1 Introduction to Hypothesis Testing

The hypothesis testing process consists of 5 Steps.

Step 4: Compute the test statistic.

Use sample data $x_1,\ldots,x_n$ and hypothesized value μ_0 to compute z (or t).

Compare test statistic z (or t) to critical value(s) $z_{\alpha/2}$ (or $t_{\alpha/2,df}$) with rule.

Step 5: Conclusion.

Make a decision, reject H_0 or not to reject H_0.

Interpret the results.

7.3 Tests with One Sample, Dichotomous Outcome

Example: Is proportion of children using dental service different from 0.86?

Step 1: Null and Alternative Hypotheses.

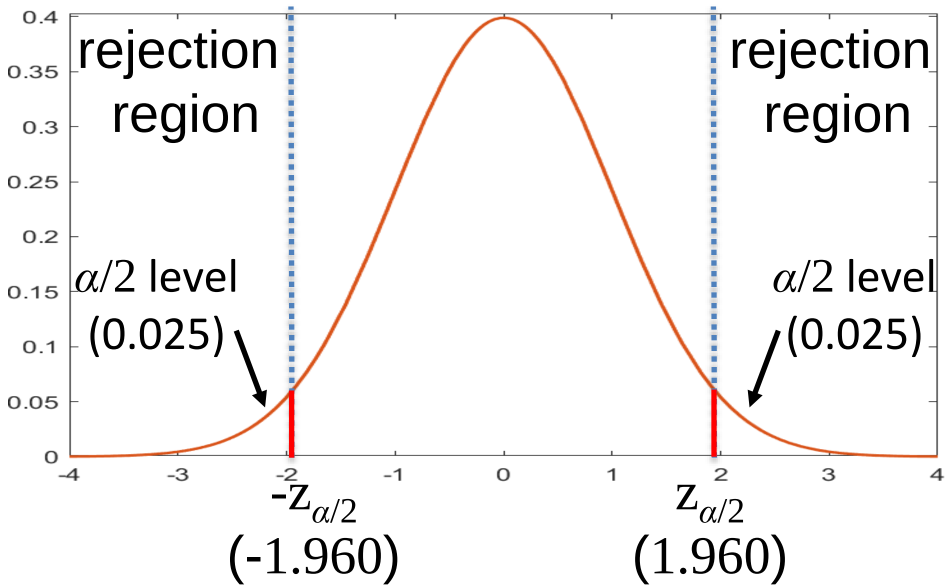
H_0: $p = 0.86$ *vs*. H_1: $p \neq 0.86$

Step 2: Test Statistic.

$$z = (\hat{p} - p_0) \big/ \sqrt{p_0(1 - p_0) / n}$$

Step 3: Decision Rule. $\alpha=0.05$

Reject H_0 if $z \leq -1.960$ or $z \geq 1.960$.



Step 4: Compute test statistic. n=125, x=64, $\hat{p} = x / n = 0.512$.

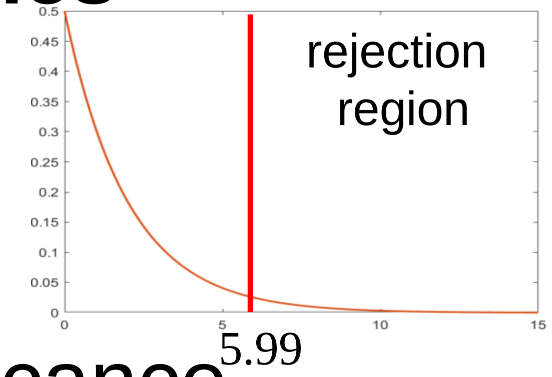
$$z = (0.512 - 0.86) \big/ \sqrt{0.86(1 - 0.86) / 125} = -11.21$$

Step 5: Conclusion

Because $z \leq -1.96$, reject and conclude proportion different from 0.86.

7.4 Tests with One Sample, Categorical and Ordinal Outcomes

Book has df=1 figure.



Example: Health Survey. $n = 470$

Step 1: Set up the hypotheses and determine the level of significance.

H_0: $p_1 = 0.60$, $p_2 = 0.25$, $p_3 = 0.15$ *vs*. H_1: H_0: false (only one pair) α=0.05

Step 2: Select the appropriate test statistic.

$$\chi^2 = \sum (O-E)^2/E \qquad df=k\text{-}1 \qquad E_i = np_{0i}$$

Step 3: Set-up the decision rule.

Reject H_0 if $\chi^2 \geq \chi^2_{0.05,2}$=5.99. ← Table 3

	No Regular Exercise	Sporadic Exercise	Regular Exercise	Total
(O)	255	125	90	470
(E)	470(0.60) = 282	470(0.25) = 117.5	470(0.15) = 70.5	470

Step 4: Compute the test statistic.

$$\chi^2 = \frac{(255-282)^2}{282} + \frac{(125-177.5)^2}{177.5} + \frac{(90-70.5)^2}{70.5} = 8.46$$

Step 5: Conclusion.

Since χ^2=8.46 $\geq \chi^2_{0.05,2}$ =5.99, reject H_0 conclude p's not what we hypothesize.

Questions?

Bring pencil/eraser, calculator, caffeinated beverage.

Will hand out exam and formula sheet/tables.