Lesson 2
Review: Topics in Statistical Inference

BIS 505b

Yale University
Department of Biostatistics

Goals for this Lesson

Addressing a Research Question

1. Estimating the parameter of interest

   - Sample statistics and confidence intervals

2. Making comparisons (testing hypotheses)

   - *One*-sample hypothesis tests
   - *Two*-sample hypothesis tests

Contents

1. Introduction
   - Motivation
   - One-Sample vs. Two-Sample
   - Probability Distributions

2. Topics in Statistical Inference
   - Central Limit Theorem
   - Confidence Intervals
   - Hypothesis Testing

Review

	Categorical	Quantitative
Descriptive Statistics	Frequency table Bar chart	Center / spread Histogram Boxplot Q-Q Plot
	Cross-tabulations	Correlation Scatterplots
Inferential Statistics	p proportion $p_1 - p_2$	μ mean $\mu_1 - \mu_2$

Review Exercise

Poll

- What are we comparing in each scenario?
  quantitative endpoint
- An investigator is interested in examining the effects of alcohol on individuals' abilities to perform physical activities. 20 randomly selected individuals consumed a prescribed amount of alcohol, and their time to complete a specific physical activity was recorded (in minutes). The investigator suspects that there may be a difference for *males* versus *females*.

- If the individual takes more than 15 minutes to complete the task, then the investigator flags this person as being "alcohol impaired." The investigator suspects that there may be a difference in *males* versus *females*.

- **Link to poll** (Lesson 02): *https://pollev.com/bis505b*

Review Exercise

Poll

- What are we comparing in each scenario?
- A study is conducted to compare two experimental treatments with respect to their effectiveness in reducing joint pain in patients with rheumatoid arthritis. 150 patients are randomly assigned to one of the two treatments. 24 hours after the assigned treatment is administered, patients are asked if they experienced an improvement in their joint pain. The investigator suspects that there may be a difference in the two groups.

binary/dichotomous

- Investigators compute the DAS28 (Disease Activity Score) for each patient before and after treatment. DAS28 combines a joint count (number swollen/tender joints), lab test (either ESR or CRP) and a VAS (Visual Analog Scale – patient questionnaire featuring a horizontal line with the words NO PAIN on the left and WORST PAIN on the right). A persistently high DAS28 has been found to be associated with increased likelihood of progressive joint damage. The investigator suspects that there may be a difference in the two groups.

- **Link to poll** (Lesson 02): *https://pollev.com/bis505b*

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Progress this Unit

1. Introduction
   - Motivation
   - One-Sample vs. Two-Sample
   - Probability Distributions

2. Topics in Statistical Inference
   - Central Limit Theorem
   - Confidence Intervals
   - Hypothesis Testing

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Motivation

- There are two main uses of statistics:

  1. Descriptive - Includes measures of location and spread, graphs and tables

  2. Inferential - Process of drawing conclusions about a population using a sample of observations

- Data are summarized by *statistics* (i.e., $\bar{x}$, $\hat{p}$), which are used to estimate population *parameters* (i.e., μ, p)

Population

μ

Sample

$\bar{x}$

(statistic)

n

(parameter)

N

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions
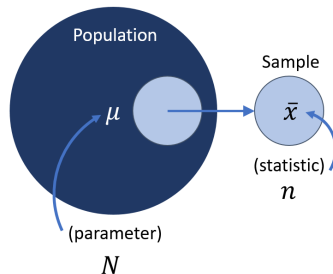
Population Parameters

Variable Type	Population Parameter
Quantitative	Population mean, μ
Categorical	Population proportion, p

- μ: The center of a Normal distribution
- p: The probability of success in a Binomial distribution

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Inferential Statistics

- The problem of inferring characteristics of a population from a sample is central to statistical inference
  - Usually interested in detecting differences
- Sampling variation produces uncertainty in inference
- Probability models provide a mechanism for making inferences about the population and for quantifying the uncertainty in inference

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Reasons for Observing a Difference

- Observe a difference, but that difference may be due to:
  1. Chance
  2. Bias
  3. Confounding } eliminate in study design

- Statistical tests are designed to rule out if the observed difference was likely due to chance

- Probability model tells us how likely is it that we would obtain the observed difference if there were no real difference in the population

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

One-Sample vs. Two-Sample Problem

Single parameter (μ, p)

$\mu_1 - \mu_2$ or $p_1 - p_2$

- One-Sample Problem - One population is of interest
  - No independent comparator or control group
  - Although there is only one group, can still be interested in a difference (e.g., difference between before and after treatment: subject serves as his/her own control, difference vs. reference or normative value) 正常值

- Two-Sample Problem - Comparing outcome of one group to another group, resulting in two groups or cohorts
  - One group generally serves as exposed or treatment group, other group serves as the unexposed or control group
  - Two samples are considered independent if those samples are selected from populations that have no effect on one another (no correlation exists between the samples)

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Exercise: Yoga Study

Poll

- Results of the 10-week pilot study showed that yoga was a fairly safe and tolerable intervention in elderly adults with chronic pain.

- At baseline, weeks 5, 10 (end of intervention), and 20 (follow-up), investigators collected data on feasibility (adherence, retention, safety), pain, anxiety, depression, fatigue, sleep disturbance, and physical function. Measured anxiety and depression using the Hospital Anxiety and Depression Scale (HADS), a brief 14-item instrument that classifies individuals as having normal, mild, moderate, or severe levels of anxiety and depressive symptoms. At the end of the intervention, anxiety significantly decreased from **5.80** (baseline) (SE = 0.90) to **4.44** (SE = 0.74) (week 10) (p = 0.014), but there were no changes in other measures.

- **Link to poll** (Lesson 02): *https://pollev.com/bis505b*

Link to article

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Exercise: Blanket Study

Poll

- Results of the randomized, controlled study show that participants using the weighted blanket for four weeks reported significantly reduced insomnia severity, better sleep maintenance, a higher daytime activity level, and reduced symptoms of fatigue, depression and anxiety compared to those using a light plastic chain blanket.

- The study involved 120 adults (68% women, 32% men) previously diagnosed with clinical insomnia and a co-occurring psychiatric disorder: major depressive disorder, bipolar disorder, attention deficit hyperactivity disorder, or generalized anxiety disorder.

- **Link to poll** (Lesson 02): *https://pollev.com/bis505b*

Link to article

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Exercise: Blanket Study

Poll

- Primary endpoint: Insomnia Severity Index (ISI), self-report rating scale assessing the severity of insomnia symptoms. Range 0-28 with higher scores indicating a more severe insomnia. [Time Frame: Change from start to 4 week use of the blanket.]

- At 4 weeks, there was a significant advantage in Insomnia Severity Index ratings of the weighted blanket intervention over the light blanket. The intervention by the weighted blanket resulted in a significantly better sleep-maintenance, a higher daytime activity level, and reduced daytime symptoms of fatigue, depression, and anxiety.

- Nearly **60%** of weighted blanket users had a positive response measured by a decrease of 50% or more in their ISI score from the baseline to the four-week endpoint, compared with **5.4%** of the control group.

- Remission, a score of seven or less on the ISI scale, was achieved by **42.2%** in the weighted blanket group, compared with **3.6%** in the control group.
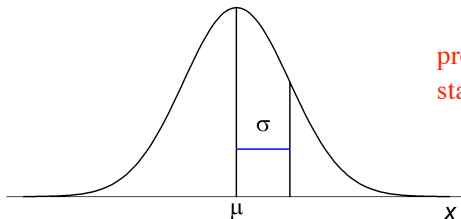
- **Link to poll** (Lesson 02): *https://pollev.com/bis505b*

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Normal Distribution

- A General Normal random variable has mean μ, standard deviation σ, $X \sim N(\mu, \sigma)$
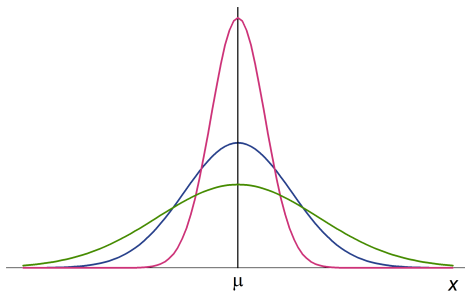
Normal Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

probability model: X ~ N
statistical test: X_bar ~ N

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Normal Distribution

- μ and σ completely determine the shape of a Normal curve

- Larger values of σ: more spread to the distribution — more variability, more noise

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Using the Normal Distribution

- Because the Normal distribution occurs so frequently, it is important to be able to calculate probabilities and percentiles associated with the Normal distribution, as well as other continuous probability distributions

- There are two main applications we will consider:

  1 Calculating probabilities for Normal random variables with specified mean μ and standard deviation σ

  - Calculating the probability a normally distributed test statistic is greater than its observed value (e.g., observe $z = 2.5$, 2-sided p-value $= 2 \times P(Z \geq 2.5)$)

  2 Finding the percentile of a standard Normal distribution

  - Calculating the $1 - \alpha/2$ percentile of Z (i.e., $z_{1-\alpha/2}$, critical value of a 2-sided test performed at the α-level based on a standard Normal test statistic)

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Probabilities from Normal Distribution

- Do not directly use $f(x)$, interested in the area under $f(x)$

- The area under the curve over an interval equals the probability that random variable X falls within that interval
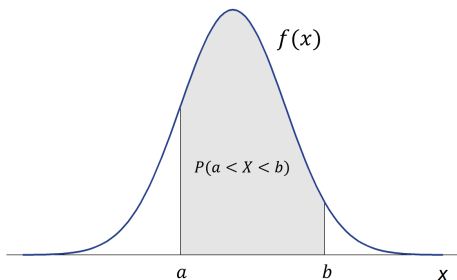
P(X ≥a): CDF cumulative density function

- Total area under the curve $= 1 = 100\%$

- $P(a < X < b) = P(X \leq b) - P(X \leq a)$

- Convert from $X \sim N(\mu, \sigma)$ to standard Normal scale, $Z \sim N(0, 1)$ using $Z = \dfrac{X - \mu}{\sigma}$

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Normal Distribution Probability Tips

- Tips to remember, based on the properties of the Normal distribution:

  1. The area under the density curve $= 1$
  2. The complement rule: $P(Z > z) = 1 - P(Z \leq z)$
  3. The Normal distribution is symmetric; the area under the density curve above z is equal to the area below $-z$
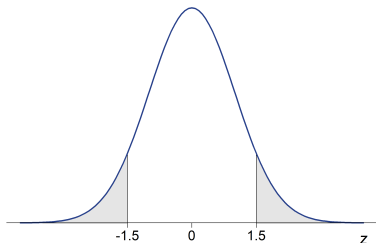
Normal Probabilities

```
# P(Z <= 1.5)
> pnorm(1.5)
[1] 0.9331928
# P(Z > 1.5) = 1 - P(Z <= 1.5)
> 1 - pnorm(1.5)
[1] 0.0668072
# P(Z <= -1.5)
> pnorm(-1.5)
[1] 0.0668072
```

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Normal Distribution Percentiles

- The kth percentile is defined as the score that holds $k\%$ of the scores below it
  - E.g., the 90th percentile in a distribution is the score that has 90% of the values below it
  - Based on α-level of test and direction of H_1, percentile determines the critical value for test statistics that follow a $N(0,1)$ distribution

Table: z-Score for Given Percentile of $N(0,1)$

Percentile	z
1	-2.326
2.5	-1.960
5	-1.645
10	-1.282
90	1.282
95	1.645
97.5	1.960

Normal Percentiles

z score = 0.01

```
# z_.01
> qnorm(.01)
[1] -2.326348

# z_.05
> qnorm(.05)
[1] -1.644854

# z_.975
> qnorm(.975)
[1] 1.959964
```

Figure: $z_{.95}$

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions
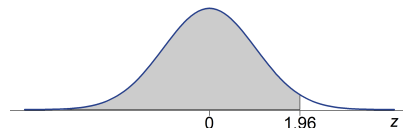
Binomial Distribution

- A Binomial random variable gives the number of successes in n trials, where p is the probability of success in each trial, $X \sim Bin(n, p)$

Binomial Probability Mass Function

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, 2, \ldots, n$$

**Figure:** Binomial(n=10, p=0.25)
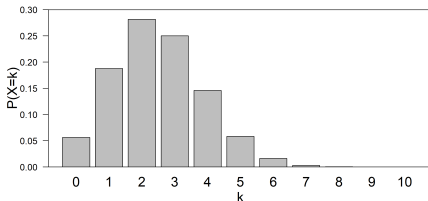
dbinom probability mass function

Binomial Probabilities

```
# P(X = 2) where X ~ bin(10, 0.25)
> dbinom(2, size = 10, prob = 0.25)
[1] 0.2815676
```

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

t Distribution

- The t distribution is used to describe the distribution of means from a Normal population distribution when n is small and σ is unknown; k is the degrees of freedom (df) of the distribution, $T \sim t_k$

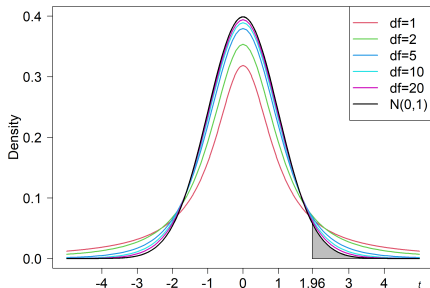- We will see some test statistics that follow a t distribution

Figure: t distributions

t(30)

0
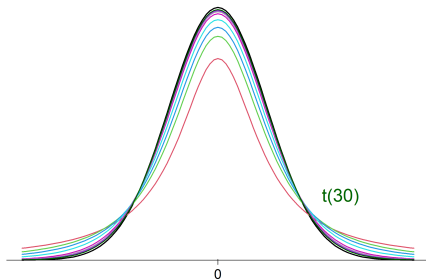
- Unimodal, symmetric about 0

- Thicker tails than $N(0, 1)$

- Reflects the extra variability introduced by the estimate s

- df depend on n: Larger samples give more reliable estimates of σ

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

t Distribution

pt

Figure: t distributions

t Probabilities

```
# P(Z > 1.96) where Z ~ N(0,1)
> 1 - pnorm(1.96)
[1] 0.0249979
# P(T > 1.96) where T ~ t(1 df)
> 1 - pt(1.96, df = 1)
[1] 0.1501714
# P(T > 1.96) where T ~ t(50 df)
> 1 - pt(1.96, df = 50)
[1] 0.02779044
```

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

t Distribution Percentiles

- Will use the percentile of the appropriate t distribution to determine critical value for test statistics that follow a t distribution

Table: 97.5th Percentile of t_{df}

df	$t_{df,.975}$
1	12.71
2	4.30
5	2.57
10	2.23
20	2.09
50	2.01
100	1.98

t Percentiles   qt

```
# z_.975
> qnorm(.975)
[1] 1.959964

# t_.975(1 df)
> qt(.975, df = 1)
[1] 12.7062

# t_.975(2 df)
> qt(.975, df = 2)
[1] 4.302653
```

Figure: t distributions

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions
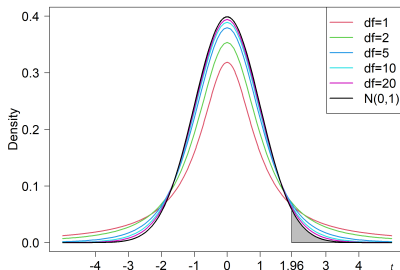
Chi-Square Distribution

- The Chi-square distribution describes the distribution of the sum of squares of k independent $N(0,1)$ random variables where k is the degrees of freedom of the distribution, $X^2 \sim \chi_k^2$

- We will see some test statistics that follow a Chi-square distribution

Figure: Chi-square distributions
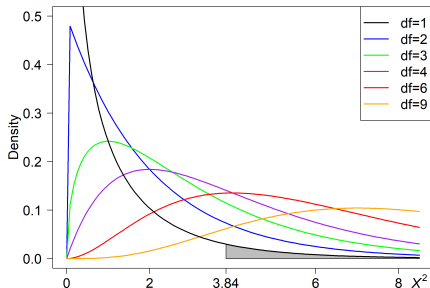
Chi-square Probabilities   pchisq

```
# P(X2 <= 3.84) where X2 ~ chisq(1 df)
> pchisq(3.84, df = 1)
[1] 0.9499565

# P(X2 > 3.84) where X2 ~ chisq(1 df)
> 1 - pchisq(3.84, df = 1)
[1] 0.05004352
```

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions

Chi-Square Distribution Percentiles

- Will use the percentile of the appropriate Chi-square distribution to determine critical value for test statistics that follow a χ^2 distribution

Table: 95th Percentile of χ^2_{df}

df	$\chi^2_{df,.95}$
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07

qchisq

Chi-square Percentiles

```
# X2_.95(1 df)
> qchisq(.95, df = 1)
[1] 3.841459

# X2_.95(2 df)
> qchisq(.95, df = 2)
[1] 5.991465
```

Figure: Chi-square distributions

Introduction
Topics in Statistical Inference

Motivation
One-Sample vs. Two-Sample
Probability Distributions
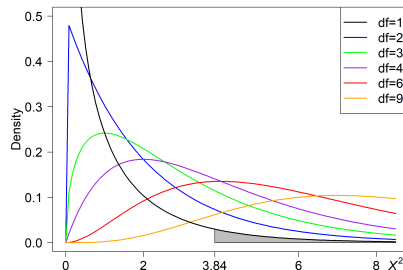
Types of R Probability Functions

- dname - Calculates the pdf/pmf of random variable X at input x, $f(x)/p(x)$; [d = density] only use in binomial, give probability mass function

- pname - Calculates the cdf of random variable X at input x (i.e., $F(x) = P(X \leq x)$); [p = probability]

- qname - Calculates the inverse cdf (i.e., the quantile/percentile q_{pr} of random variable X where $P(X \leq q_{pr}) = pr$); [q = quantile]

- rname - Generates a random value from the indicated distribution; [r = random]

R Code and Output

```
> pnorm(1.96)            # P(Z<=1.96)
[1] 0.9750021
> qnorm(.975)            # z_.975
[1] 1.959964
> dbinom(2, size = 5, prob = 0.25) # P(X=2)
[1] 0.2636719
> pbinom(2, size = 5, prob = 0.25) # P(X<=2)
[1] 0.8964844
> pt(12.71, df = 1)      # P(T<=12.71)
[1] 0.9750074
> qt(.975, df = 1)       # t_.975(1 df)
[1] 12.7062
> pchisq(3.84, df = 1)   # P(X2<=3.84)
[1] 0.9499565
> qchisq(.95, df = 1)    # X2_.95(1 df)
[1] 3.841459
```

Progress this Unit

1. Introduction
   - Motivation
   - One-Sample vs. Two-Sample
   - Probability Distributions

2. Topics in Statistical Inference
   - Central Limit Theorem
   - Confidence Intervals
   - Hypothesis Testing

p-value
critical value

Statistical Inference

- There are two broad areas of statistical inference

  1 Estimation
    - The population parameter (μ, p, $\mu_1 - \mu_2$, $p_1 - p_2$) is unknown
    - A random sample is drawn from the population of interest and sample statistics are used to estimate the parameter and an interval that is likely to contain the parameter

  2 Hypothesis Testing
    - An explicit hypothesis is stated about the population parameter
    - A random sample is drawn from the population of interest and sample statistics are used to determine whether to support or reject the hypothesis about the parameter

Estimation

- The parameters we are trying to estimate are the population mean μ and the population proportion p

- In the two-sample case, we are interested in comparing parameters in two populations (i.e., $\mu_1 - \mu_2$, $p_1 - p_2$)

- Two methods of estimation:

    1. **Point Estimation** - Use the data to calculate a single number to estimate the parameter of interest, *sample statistic*

    2. **Interval Estimation** - Provides a reasonable range of values that is intended to contain the parameter with a certain degree of confidence, *confidence interval*

Point Estimate

- Point estimate (sample statistic) - Single value used to estimate a population parameter from a sample of observations

Probability Distribution	Population Parameter	Point Estimate
Normal	μ	$\hat{\mu} = \bar{x} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i$
Binomial	p	$\hat{p} = \dfrac{x}{n}$

- $\mu_1 - \mu_2$ is estimated by $\bar{x}_1 - \bar{x}_2$; $p_1 - p_2$ is estimated by $\hat{p}_1 - \hat{p}_2$
- For the Binomial case, $x =$ Observed number of successes in n trials

Sampling Variability

- Statistics are random variables that can vary from sample to sample

- For example, the sample means, $\bar{x}$, tend to cluster around the mean of the population, μ; there is less variability among the sample means than among the individual observations

- Sampling variability of a statistic refers to how much the statistic varies from sample to sample (measured by the spread of the sampling distribution, or the statistic's standard error)

- Quantifying how sample statistics vary provides a way to estimate the margin of error associated with the point estimate

CLT: Example

- Consider the distribution of age at death in U.S. population in 2018 (2.8 million deaths)

Relative Frequency Histogram of Age at Death, 2018
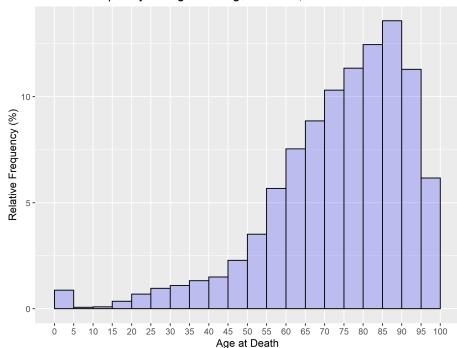
R Code and Output

```
> mean(dth)     # mu
[1] 73.29689
> median(dth)
[1] 77
> sd(dth)       # sigma
[1] 18.04073
```

- If we take a sample of size $n = 50$ from this population and created a histogram, what shape would we expect the histogram of death ages in that sample to have?

CLT: Example

Figure: Simple Random Sample, $n = 50$

Mean = 72.9 ; SD = 18.6
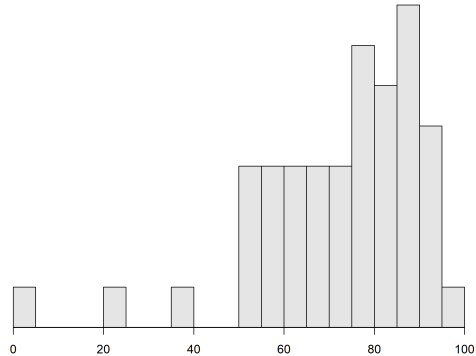
CLT: Example

- Selecting a random sample of size 50 from the population of ages at time of death for the U.S. population

- Sample histograms show shapes similar to the population distribution

Sample of Size 50	$\bar{x}$	s
1	72.9	18.6
2	71.7	17.2
3	73.6	20.2
4	77.7	17.9

- Note that the four random samples are not identical: sampling variability

CLT: Example

Figure: Simple Random Sample, $n = 100$

Mean = 72.5 ; SD = 17.8
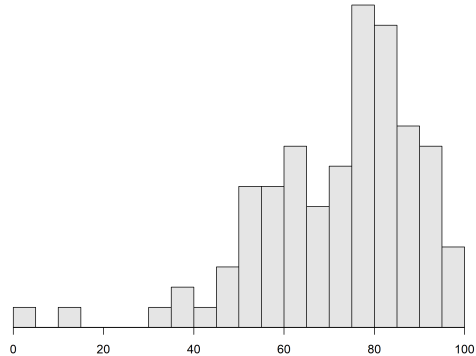
CLT: Example

Sample of Size 100	$\bar{x}$	s
1	72.5	17.8
2	70.0	21.4
3	70.5	17.0
4	74.6	19.8

- As the sample size increases, the amount of sampling variability (i.e., the variability in the sample means themselves) decreases
    - Samples of size 50: $\bar{x}$ ranged from 71.7 to 77.7, range $= 6.0$
    - Samples of size 100: $\bar{x}$ ranged from 70 to 74.6, range $= 4.6$

CLT: Example

Figure: Histogram of 10,000 sample means from samples of size 50

Histogram of Sample Means from Samples of Size n=50 and N(73.3, 2.55)
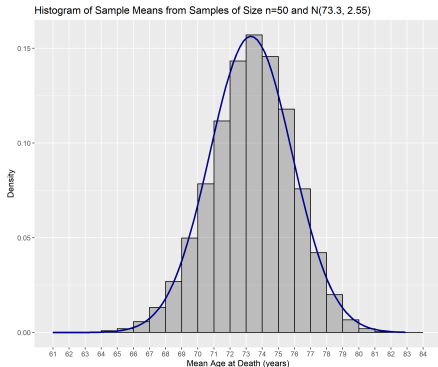
- Although the distribution of survival times is heavily skewed, if the sample size is large enough, the distribution of sample means is approximately Normal

- Sampling distribution:
  $\bar{X} \sim N(\mu = 73.297, \frac{\sigma}{\sqrt{n}} = \frac{18.04}{\sqrt{50}} = 2.55)$

central limit theorem

R Code and Output

```
> mean(dth)      # population of death times
[1] 73.29689         # true mu
> sd(dth)/sqrt(50)
[1] 2.551344         # true sigma/sqrt(50)
> mean(means50) # vector of sample means
[1] 73.30163
> sd(means50)
[1] 2.549788
```

Sampling Variability

Probability Distribution	Population Parameter	Point Estimate	Estimate of Sampling Variation[†]
Normal	μ	$\hat{\mu} = \bar{x} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i$	$s_{\bar{x}}^2 = \dfrac{s^2}{n}$
	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}$
Binomial	p	$\hat{p} = \dfrac{x}{n}$	$s_{\hat{p}}^2 = \dfrac{\hat{p}(1-\hat{p})}{n}$
	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}$

- $Var(X - Y) = Var(X) + Var(Y) - 2 \times Cov(X, Y)$
- There is less variability among the sample means than among the individual observations
- [†]$\sqrt{\text{Sampling variation}}$ = Standard error
- Normality of sampling distribution: Normal ($n \geq 30$); Binomial (at least 5 successes and 5 failures)

Interval Estimate

- One way to answer a research question is to provide an estimate of the parameter of interest (point estimate)
  - A statistic from a representative sample of the population might provide a good guess or estimate of the parameter, but a sample is unlikely to provide a statistic that is exactly equal to the population parameter
- If the statistic (point estimate) is made into an interval (interval estimate), the interval has a larger chance of including the unknown parameter

Confidence Interval

$100(1 - \alpha)\%$ confidence interval is a range of values that is likely to cover the true parameter; interval contains the parameter with probability $(1 - \alpha)$

Normal Distribution: Empirical Rule

- Use the properties of the sampling distribution to construct an interval that is likely to contain the true parameter with high probability

- Since the sampling distribution is Normal, the Empirical Rule can be used to determine the proportion of observations (statistics) that are 1, 2, or 3 standard errors from the mean μ
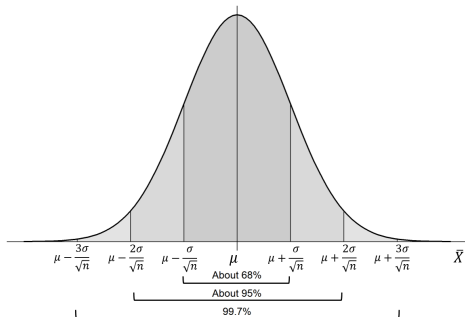

- $\approx 68\%$ of all sample means are within 1 standard error of the true parameter (μ)
- $\approx 95\%$ within 2 standard errors
- 99.7% within 3 standard errors
- If two standard errors are $+$ and - to every possible $\bar{x}$, $\approx 95\%$ of them would result in interval estimates that would include the true population mean

Interpretation of the Confidence Interval

- Suppose the true distribution of age of death has $\mu = 73.3$ and $\sigma = 18.0$

Figure: CIs for mean age of death from 100 random samples

Sample / Mean Age of Death

68.3   73.3   78.3
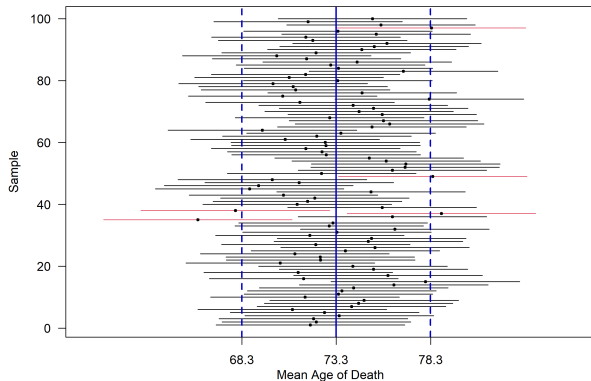
- Taking 100 simple random samples of size $n = 50$, we expect 5 of the samples to give a 95% confidence interval that does not capture μ

- Interpretation: There is 95% probability that the confidence interval contains μ

- 5 of the 100 samples had $\bar{x}$ outside the range $(68.3, 78.3)$ (note: $1.96 \times 2.55 = 5$)

Controlling the Width of the Confidence Interval

- The basic structure of a confidence interval is:

$$\text{point estimate} \pm \underbrace{\text{multiplier} \times \text{standard error}}_{m}$$

- Margin of error $(m) = $ multiplier $\times$ standard error

- The width of the confidence interval is affected by:

  1. Multiplier $(z_{1-\frac{\alpha}{2}}$ or $t_{df,1-\frac{\alpha}{2}})$ (i.e., the level of confidence, $1-\alpha$)
  2. Standard error

Confidence Level

point estimate $\pm$ multiplier $\times$ standard error

- Increasing the confidence level $(1 - \alpha)$ will increase the multiplier and the width of CI
  - $\downarrow \alpha, \quad \uparrow z_{1-\frac{\alpha}{2}}$ (or $t_{df, 1-\frac{\alpha}{2}}$, as appropriate)

trade-off:
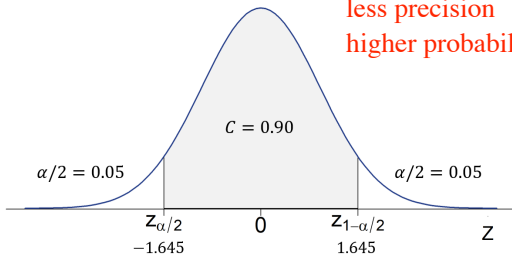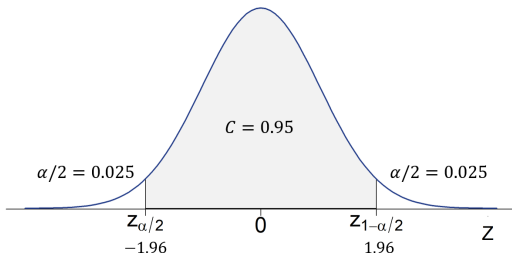less precision
higher probability of contain true value

Table: CI multipliers, Normal

Confidence Level (%)	$z_{1-\frac{\alpha}{2}}$	α
99.99	3.819	0.0001
99.9	3.291	0.001
99	2.576	0.01
95	1.960	0.05
90	1.645	0.10

Confidence Level

point estimate $\pm$ multiplier $\times$ standard error

- Increasing the confidence level $(1 - \alpha)$ will increase the multiplier and the width of CI
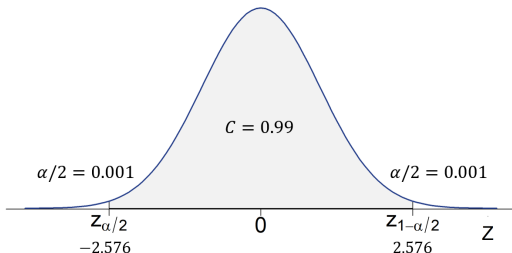   - $\downarrow \alpha, \quad \uparrow z_{1-\frac{\alpha}{2}}$ (or $t_{df, 1-\frac{\alpha}{2}}$, as appropriate)

$\alpha/2 = 0.025$ $C = 0.95$ $\alpha/2 = 0.025$

$z_{\alpha/2}$ 0 $z_{1-\alpha/2}$ z
-1.96 1.96

Table: CI multipliers, Normal

Confidence Level (%)	$z_{1-\frac{\alpha}{2}}$	α
99.99	3.819	0.0001
99.9	3.291	0.001
99	2.576	0.01
95	1.960	0.05
90	1.645	0.10

Confidence Level

point estimate $\pm$ multiplier $\times$ standard error

- Increasing the confidence level $(1 - \alpha)$ will increase the multiplier and the width of CI
  - $\downarrow \alpha$, $\quad \uparrow z_{1-\frac{\alpha}{2}}$ (or $t_{df,1-\frac{\alpha}{2}}$, as appropriate)
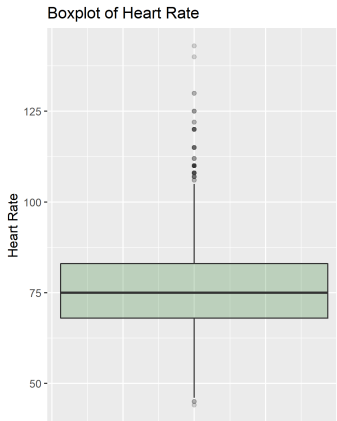

$C = 0.99$

$\alpha/2 = 0.001$

$\alpha/2 = 0.001$

$z_{\alpha/2}$
-2.576

0

$z_{1-\alpha/2}$
2.576

z

Table: CI multipliers, Normal

Confidence Level (%)	$z_{1-\frac{\alpha}{2}}$	α
99.99	3.819	0.0001
99.9	3.291	0.001
99	2.576	0.01
95	1.960	0.05
90	1.645	0.10

CI for μ
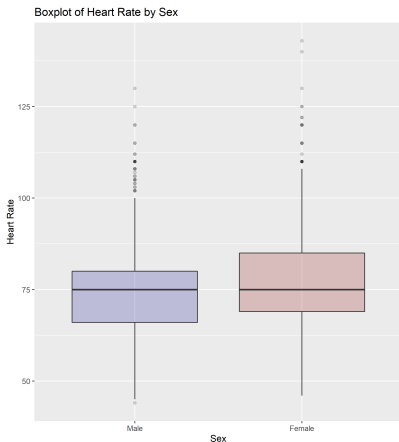

Boxplot of Heart Rate

R Code and Output, CI for μ

```
# One-sample CI for mu
> mean(fhs$HEARTRTE, na.rm = TRUE)
[1] 75.89104

> t.test(fhs$HEARTRTE, conf.level = 0.95)$conf.int
[1] 75.53435 76.24774
attr(,"conf.level")
[1] 0.95
```

- 95% CI for mean heart rate μ: (75.53, 76.25)

CI for $\mu_1 - \mu_2$

Boxplot of Heart Rate by Sex

R Code and Output, CI for $\mu_1 - \mu_2$

```
# Two-sample CI for mu1 - mu2
> aggregate(x = list(meanhr = fhs$HEARTRTE),
            by = list(sex = fhs$SEX_factor),
            FUN = mean, na.rm = TRUE)
     sex   meanhr
1   Male 74.39732
2 Female 77.05663
# analysis variable ~ grouping variable
> t.test(HEARTRTE ~ SEX_factor, data - fhs,
        conf.level = 0.95)$conf.int
[1] -3.372185 -1.946421
...
```

- Point estimate of **difference** in mean HR in males vs. females: -2.66
- 95% CI $\mu_m - \mu_f$: (-3.37, -1.95)

CI for p

Relative Frequency Bar Plot of CVD Status
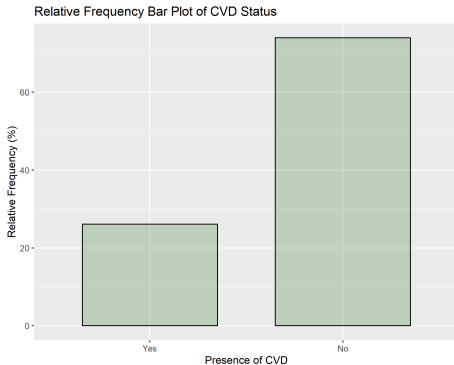
- 95% CI for prevalence of CVD p: (0.248, 0.274)

R Code and Output, CI for p

```
# One-sample CI for p
> tab1p <- table(fhs$CVD_factor, dnn = "CVD")
> tab1p
CVD
 Yes   No
1157 3277
> prop.table(tab1p)        relative frequency
CVD
      Yes        No
0.2609382 0.7390618

> prop.test(tab1p, conf.level = 0.95)$conf.int
[1] 0.2481124 0.2741815
attr(,"conf.level")
[1] 0.95
```

CI for $p_1 - p_2$

Relative Frequency Barplot of CVD within Sex
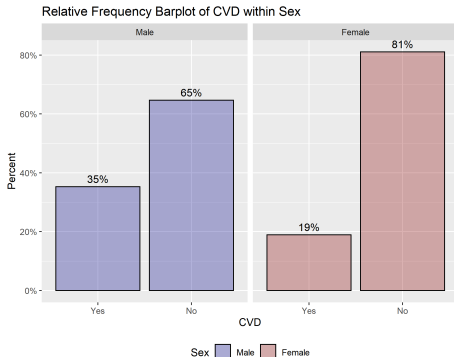
- Point estimate of risk difference of CVD in M vs. F: 0.164
- 95% CI for risk difference of CVD $p_m - p_f$: (0.137, 0.190)

R Code and Output, CI for $p_1 - p_2$

```
# Two-sample CI for p1-p2
> tab2p <- table(fhs$SEX_factor, fhs$CVD_factor,
                 dnn = c("Sex", "CVD"))
> tab2p
        CVD
Sex       Yes    No
  Male    686  1258
  Female  471  2019
> prop.table(tab2p, margin = 1) # P(D|M), P(D|F)
        CVD
Sex           Yes        No
  Male    0.3528807 0.6471193
  Female  0.1891566 0.8108434
> prop.test(tab2p, conf.level = 0.95)$conf.int
[1] 0.1370388 0.1904093
attr(,"conf.level")
[1] 0.95
```

CI for Odds Ratio

- The odds ratio is limited at the lower end (non-negative) but not at the upper end, giving a skewed distribution
- The log odds ratio can take any value and $\dot\sim N \rightarrow$ construct CI for $\ln \text{OR}$

  approximately normally distributed

	Disease	No Disease	
Exposed	a	b	$a + b$
Unexposed	c	d	$c + d$
	$a + c$	$b + d$	N

Measure	Estimate	$100(1 - \alpha)\%$ CI for OR*
Odds Ratio	$\hat{\text{OR}} = \dfrac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_2}{1 - \hat{p}_2}} = \dfrac{ad}{bc}$	$\exp\left(\ln \hat{\text{OR}} \pm z_{1 - \frac{\alpha}{2}} \sqrt{\dfrac{1}{a} + \dfrac{1}{b} + \dfrac{1}{c} + \dfrac{1}{d}} \right)$

*All cell frequencies $(a, b, c, d) \geq 5$

CI for Odds Ratio

R Code and Output, CI for OR

```
> library(epiR)
> epi.2by2(tab2p, method = "cohort.count", units = 1)
           Outcome +    Outcome -     Total       Inc risk *     Odds
Exposed +        686         1258      1944           0.353       0.545
Exposed -        471         2019      2490           0.189       0.233
Total           1157         3277      4434           0.261       0.353

Point estimates and 95% CIs:
-----------------------------------------------------------------
Inc risk ratio                      1.87 (1.69, 2.06)
Odds ratio                          2.34 (2.04, 2.68)
Attrib risk *                       0.16 (0.14, 0.19)
Attrib risk in population *         0.07 (0.05, 0.09)
Attrib fraction in exposed (%)      46.40 (40.69, 51.55)
Attrib fraction in population (%)   27.51 (22.97, 31.78)
-----------------------------------------------------------------
 Test that OR = 1: chi2(1) = 151.742 Pr>chi2 = <0.001
 Wald confidence limits      CI: confidence interval      * Outcomes per population unit
```

Public Health Application: Prevalence Studies

Public Health Application

Prevalence studies use data from a sample of respondents to estimate the frequency and distribution of different conditions. Often used to assess the burden of disease in a population to assess the need for health services. Prevalence studies may also compare the prevalence of disease in different populations or examine trends in disease prevalence or severity over time. The focus could be a narrow one (e.g., a particular clinical setting) or a broad one (e.g., an entire nation).

A valid study uses a sample that accurately represents a defined target population. It is important to report confidence intervals for the unknown parameters to give a range of values that is likely to contain the true parameter (p) with high probability.

Link to article, "Guidelines for evaluating prevalence studies"

Public Health Application: Obesity in Children and Adolescents

Public Health Application

Analysis conducted to provide the most recent data on the prevalence of obesity and severe obesity among United States children and adolescents aged 2 to 19 years.

Table: Prevalence of Overweight and Obesity Among Children and Adolescents, 2015–2016

Overweight	Class I Obesity	Class II Obesity	Class III Obesity
35.1 (31.9, 38.4)	18.5 (15.8, 21.2)	6.0 (4.3, 7.6)	1.9 (1.0, 2.9)

There are few differences in the prevalence of overweight and all classes of obesity since the last NHANES cycle, 2013–2014. Non-Hispanic African American and Hispanic children had higher prevalence of overweight and all classes of obesity compared with other races. Asian American children had markedly lower prevalence of overweight and all classes of obesity.

Link to article

Hypothesis Testing

- Hypothesis testing is the process used to evaluate the strength of evidence from a sample and provides a framework for drawing conclusions about the population

- Method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from which the sample was drawn

- The investigator formulates a specific hypothesis, evaluates data from the sample, and uses those data to decide whether they support the specific hypothesis

    - Determine if the data provide evidence supporting rejection of the null hypothesis H_0 which is generally one of no difference in favor of the alternative hypothesis H_1 which claims there is a difference

Using the Sampling Distribution

- Confidence intervals use the sampling distribution to specify an interval that is likely to contain the true parameter

- Hypothesis tests use the sampling distribution to determine whether there is evidence against a null parameter

  - The sampling distribution used in hypothesis testing is the sampling distribution that *would* exist *if* the null hypothesis were true

Null and Alternative Hypotheses

- Null hypothesis (H_0) is usually a statement of no effect or no difference

Table: Null Hypotheses

One-Sample Case	Two-Sample Case
$\mu = \mu_0$	$\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$
$p = p_0$	$p_1 = p_2$ or $p_1 - p_2 = 0$

- Alternative hypothesis $(H_1$ or $H_a)$ or the research hypothesis states there is a difference between the distribution of the observed data and the null-hypothesized distribution
  - Two-sided alternative $\quad \mu = \mu_0$
  - One-sided alternative

Data Present Evidence Against H_0

- Example: Is the mean serum cholesterol level in the population of hypertensive smokers different from 220? Serum cholesterol is assumed to be normally distributed with $\sigma = 40$.

$$H_0 : \mu = 220$$
$$H_1 : \mu \neq 220$$

- Main idea: If there is evidence that the sample could **not** have come from a population with mean $\mu_0 = 220$, we reject H_0 and conclude H_1

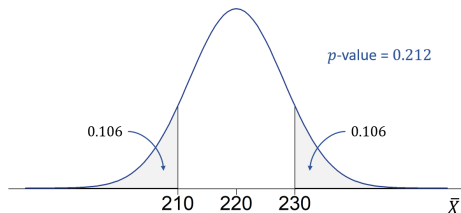**Figure:** Distribution of $\bar{X}$ under H_0; $\bar{X} \sim N(200, 8)$

0.95

204   220   236 $\bar{X}$
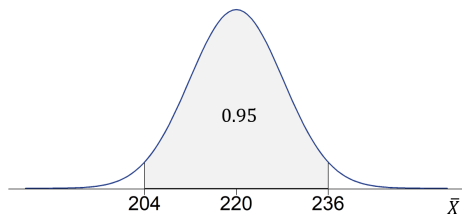
- Compare the mean of the sample, $\bar{x}$, ($n = 25$) to the postulated mean $\mu_0 = 220$

- Is the difference observed too large to be attributed to chance alone?

- Note: $\frac{\sigma}{\sqrt{n}} = \frac{40}{5} = 8$

When to Reject H_0

- If the probability of obtaining an $\bar{X}$ as extreme or more extreme than $\bar{x}$ assuming H_0 is true (p-value) is sufficiently small, reject H_0 and conclude H_1

- p-value $\leq \alpha$ (e.g., $\alpha = 0.05$)

  - Implies that 5% of the time, will incorrectly reject H_0

  - α - Significance level of the test, or the probability of committing a Type I error
    $P(\text{reject } H_0 | H_0 \text{ true})$

When to Reject H_0: Example

$$H_0 : \mu = 220$$
$$H_1 : \mu \neq 220$$

Figure: Distribution of $\bar{X}$ under H_0; $\bar{X} \sim N(220, 8)$

p-value = 0.212   > 0.05, fail to reject H0

0.106   0.106

210   220   230 $\bar{X}$

- $\bar{x} = 210$

- *p*-value: Probability of obtaining an $\bar{X}$ as extreme or more extreme than $\bar{x}$ in the direction of H_1 assuming H_0 is true

When to Reject H_0: Example

Figure: Distribution of $\bar{X}$ under H_0; $\bar{X} \sim N(220, 8)$
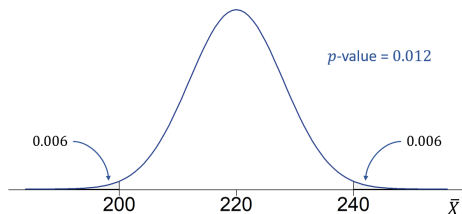
$$H_0 : \mu = 220$$
$$H_1 : \mu \neq 220$$

- $\bar{x} = 240$

- p-value: Probability of obtaining an $\bar{X}$ as extreme or more extreme than $\bar{x}$ in the direction of H_1 assuming H_0 is true

Central Limit Theorem in Hypothesis Testing

- Instead of determining the critical values for $\bar{X}$ that would be specific to each application, we apply the Central Limit Theorem and work on a standard Normal (Z) scale

- Assuming $H_0 : \mu = \mu_0$ is true, X has mean μ_0. If the standard deviation σ is known,

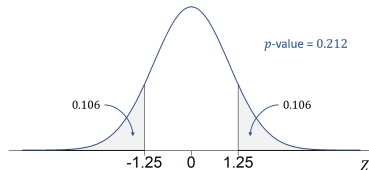$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \; \dot\sim \; N(0, 1) \qquad \text{if } n \text{ is sufficiently large}$$
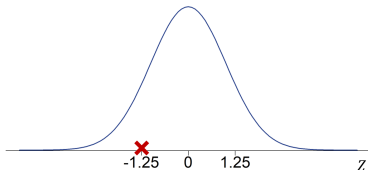
- When σ is not known, it is replaced by the estimate $\hat{\sigma} = s$, and Z is replaced by:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

  - When $n \geq 30$, the Normal approximation may be used

Central Limit Theorem in Hypothesis Testing: Example

- Under H_0, $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{\bar{X} - 220}{40/\sqrt{n}} \ \dot\sim \ N(0,1)$
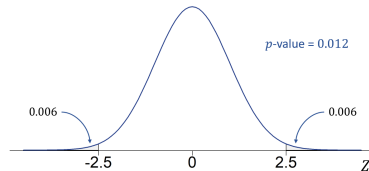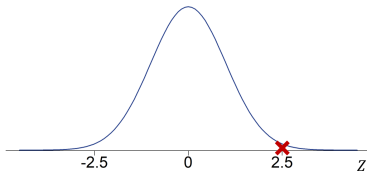

- $\bar{x} = 210$, $n = 25$

- $z = \dfrac{\bar{x} - 220}{\sigma/\sqrt{n}} = \dfrac{210 - 220}{40/\sqrt{25}} = -1.25$

- $p = 2 \times P(Z \geq |z|) = 2 \times 0.106 = 0.212$

R Code and Output

```
> zstat = (210 - 220)/(40/sqrt(25))
> zstat
[1] -1.25
> pvalue = 2*(1 - pnorm(abs(zstat)))
> pvalue
[1] 0.2112995
```

Central Limit Theorem in Hypothesis Testing: Example

- Under H_0, $Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{\bar{X} - 220}{40/\sqrt{n}} \; \dot\sim \; N(0,1)$

- $\bar{x} = 240$, $n = 25$

- $z = \dfrac{\bar{x} - 220}{\sigma/\sqrt{n}} = \dfrac{240 - 220}{40/\sqrt{25}} = 2.5$

- $p = 2 \times P(Z \geq |z|) = 2 \times 0.006 = 0.012$

R Code and Output

```
> zstat = (240 - 220)/(40/sqrt(25))
> zstat
[1] 2.5
> pvalue = 2*(1 - pnorm(abs(zstat)))
> pvalue
[1] 0.01241933
```

One-Sample t-Test

- σ generally unknown, so estimate by s

- Under H_0,

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\bar{X} - 220}{s/\sqrt{n}} \sim t_{n-1}$$

- $\bar{x} = 239.48$, $s = 40.14$, $n = 25$

- $t = \dfrac{\bar{x} - 220}{s/\sqrt{n}} = \dfrac{239.48 - 220}{40.14/\sqrt{25}} = 2.43$

- $p = 2 \times P(T \geq |t|) = 2 \times 0.0116 = 0.023$

R Code and Output

```
> mn = mean(fhshms25$TOTCHOL, na.rm = TRUE); mn
[1] 239.48
> sdev = sd(fhshms25$TOTCHOL, na.rm = TRUE); sdev
[1] 40.13947
> n = sum(!is.na(fhshms25$TOTCHOL)); n
[1] 25
> tstat = (mn - 220)/(sdev/sqrt(n)); tstat
[1] 2.42654        test statistics
> qt(.975, df = n - 1)        # t_.975(24 df)
[1] 2.063899        critical value
> pval.t = 2*(1 - pt(abs(tstat), df = n - 1)); pval.t
[1] 0.0231233

# Normal (if sigma had been known and equal to s)
> qnorm(.975)                 # z_.975
[1] 1.959964
> pval.z = 2*(1 - pnorm(abs(tstat))); pval.z
[1] 0.01524359
```

One-Sample t-Test

R Code and Output

```
# One-sample t-test
> t.test(fhshms25$TOTCHOL,
         mu = 220,                      # mu0
         alternative = "two.sided")    # direction

    One Sample t-test

data:  fhshms25$TOTCHOL
t = 2.4265, df = 24, p-value = 0.02312
alternative hypothesis: true mean is not equal to 220
95 percent confidence interval:
 222.9112 256.0488
sample estimates:
mean of x
   239.48
```

- $t = \dfrac{\bar{x} - 220}{s/\sqrt{n}} = \dfrac{239.48 - 220}{40.14/\sqrt{25}} = 2.43$ $p = 2 \times P(T \geq |t|) = 2 \times 0.0116 = 0.023$

Equivalence between CI and 2-Sided Hypothesis Test for μ

μ_0	Conclusion
$\mu_0 \in 100(1-\alpha)\%$ CI for μ	Fail to reject H_0 at α-level
$\mu_0 \notin 100(1-\alpha)\%$ CI for μ	Reject H_0 at α-level

- For a 2-tailed hypothesis test of $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, if the confidence interval does not contain the value specified in H_0 (μ_0), then we reject H_0

- Example: $H_0 : \mu = 220$ vs. $H_1 : \mu \neq 220$

  - $n = 25$, $\dfrac{s}{\sqrt{n}} = \dfrac{40.14}{\sqrt{25}} = 8.03$ $\hspace{2cm}$ 95% CI: $\bar{x} \pm t_{24;.975} \dfrac{s}{\sqrt{n}}$

 | μ_0 | $\bar{x}$ | 95% CI | Conclusion |
  |---------|-----------|--------|------------|
  | 220 | 239.48 mg/100 ml | (222.91, 256.05) | Reject H_0 at $\alpha = 0.05$-level |

Steps in Hypothesis Testing

Steps in Hypothesis Testing

1. State H_0 and H_1

2. Specify the significance level (e.g., $\alpha = 0.05$)

3. Compute or report the appropriate test statistic

4. Generate the decision rule

   - e.g., Given $\alpha = 0.05$ and a two-tailed test is performed, Reject H_0 if $|z| \geq z_{1-\frac{\alpha}{2}} = z_{.975} = z^* = 1.96$

5. Draw a conclusion about H_0 by:

   - Comparing test statistic to decision rule and/or computing p-value
     - Reject H_0 if test statistic in rejection region $\equiv p \leq \alpha$
     - Fail to reject H_0 if test statistic not in rejection region $\equiv p > \alpha$

6. state conclusion in the context of the problem

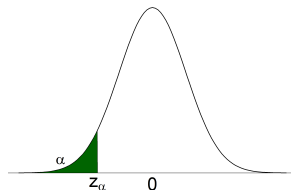## Examples of Decision Rules: Z-test

Two-sided test
$H_0 : \mu = \mu_0$ vs.
$H_1 : \mu \neq \mu_0$

Lower-tailed test
$H_0 : \mu = \mu_0$ vs.
$H_1 : \mu < \mu_0$

Upper-tailed test
$H_0 : \mu = \mu_0$ vs.
$H_1 : \mu > \mu_0$
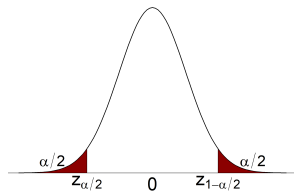
- Reject H_0 if $z \leq z_{\frac{\alpha}{2}}$
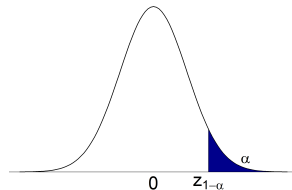 or if $z \geq z_{1-\frac{\alpha}{2}}$
  $p = 2 \times P(Z \geq |z|)$

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

  $p = P(Z \leq z)$

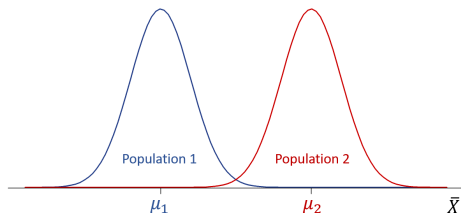- Reject H_0 if $z \geq z_{1-\alpha}$

  $p = P(Z \geq z)$

- In each case, there is probability α of rejecting H_0 when H_0 is true
- If z falls into the rejection region, reject H_0

Two-Sample t-Test

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Figure: Distribution of $\bar{X}_1$ and $\bar{X}_2$

- In a two-sample setting, the goal is to compare the mean in two independent populations, μ_1 and μ_2

- Under H_0, there is no difference in the mean response in the two populations, $\mu_1 - \mu_2 = 0$

Two-Sample t-Test

- Question: Is mean heart rate significantly different in males and females?

- μ: Mean heart rate, $H_0 : \mu_m = \mu_f$

Table 1: Comparing Male and Female Heart Rate in FHS Data

Sex	Male (N=1944)	Female (N=2490)	p value
Heart Rate			< 0.001
Missing	1	0	
Mean (SD)	74.40 (11.90)	77.06 (12.15)	
Mean (95% CI)	74.40 (73.87, 74.93)	77.06 (76.58, 77.53)	

Two-Sample t-Test

R Code and Output

```
# Summary statistics of heart rate by group
> aggregate(x = list(hr = fhs$HEARTRTE),
            by = list(sex = fhs$SEX_factor),
            FUN = function(x) c(mn = mean(x, na.rm = TRUE), sdev = sd(x, na.rm = TRUE)))
     sex   hr.mn   hr.sdev
1   Male 74.39732 11.90373
2 Female 77.05663 12.15037
# analysis variable ~ grouping variable, var.equal = FALSE by default (unpooled t-test)
> t.test(HEARTRTE ~ SEX_factor, data = fhs, alternative = "two.sided")
     Welch Two Sample t-test
data:  HEARTRTE by SEX_factor
t = -7.3135, df = 4211.4, p-value = 3.1e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.372185 -1.946421
sample estimates:
  mean in group Male mean in group Female
            74.39732             77.05663
```

Equivalence between CI and 2-Sided Hypothesis Test for $\mu_1 - \mu_2$

- Recall, under H_0, $\mu_1 = \mu_2$, or $\mu_1 - \mu_2 = 0$

$\mu_1 - \mu_2$	Conclusion
$0 \in 100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$	Fail to reject H_0 at α-level
$0 \notin 100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$	Reject H_0 at α-level

- For a 2-tailed hypothesis test of $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$, if the confidence interval does not contain the value specified in H_0 (0), then we reject H_0

- Example: $H_0 : \mu_m - \mu_f = 0$ vs. $H_1 : \mu_m - \mu_f \neq 0$

$\mu_m - \mu_f$	$\bar{x}_m - \bar{x}_f$	95% CI	Conclusion
0	-2.66	(-3.37, -1.95)	Reject H_0 at $\alpha = 0.05$-level

One-Sample Test for Proportions

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

- **Question:** Is the prevalence of CVD in the Framingham population different from the current prevalence of CVD in the U.S., $p_0 = 0.48$?

- One-sample test of binomial proportions can be based on the $N(0,1)$ or χ_1^2 distribution[*]

Link to article
[*] $\min(np_0; n(1 - p_0)) \geq 5$

R Code and Output

```
> tab1p <- table(fhs$CVD_factor, dnn = "CVD")
> prop.table(tab1p)
CVD
      Yes        No
0.2609382 0.7390618
> prop.test(tab1p, p = 0.48,            # p0
                    alternative = "two.sided")  # direction

1-sample proportions test with continuity correction

data:  tab1p, null probability 0.48
X-squared = 851.6, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.48
95 percent confidence interval:
 0.2481124 0.2741815
sample estimates:
        p
0.2609382
```

Two-Sample Test for Proportions

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

- **Question:** Is the prevalence of CVD in males different from females? Is there an association between CVD and sex?

- Chi-square test commonly used and can be used with larger tables ($> 2 \times 2$)

Table 2: Comparing Male and Female CVD in FHS Data

Sex	Male (N=1944)	Female (N=2490)	p value
Presence of CVD			< 0.001
Yes	686 (35.3%)	471 (18.9%)	
No	1258 (64.7%)	2019 (81.1%)	

R Code and Output

```
> tab2p <- table(fhs$SEX_factor, fhs$CVD_factor,
                 dnn = c("Sex", "CVD"))
> prop.table(tab2p, margin = 1) # row proportions
        CVD
Sex          Yes        No
  Male   0.3528807 0.6471193
  Female 0.1891566 0.8108434

> prop.test(tab2p, alternative = "two.sided")

2-sample test for equality of proportions with cont.corr.

data:  tab2p
X-squared = 150.89, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1370388 0.1904093  ...

> chisq.test(tab2p)              # Chi-square test
    Pearson's Chi-squared test with Yates' cont.corr.
data:  tab2p
X-squared = 150.89, df = 1, p-value < 2.2e-16

> 1 - pchisq(150.89, df = 1)   # p-value in upper tail only
[1] 0
```

Lesson Summary

- The review lessons provided an overview of the basic statistical concepts encountered in public health research:
    - Describing data
    - Estimating the parameter
    - Testing the hypothesis

- Organizing the methods according to the research question (or type of variable measured) simplifies choosing the appropriate statistical method

Flow Chart of Descriptive and Inferential Statistics