Lecture 05 - Standard Scores and the Normal Distribution

Samuel E. DeWitt, Ph.D.

Organization

Today we will be talking about a few things, chief among them are:

1) The Normal Distribution
2) Sample, population, and sampling distributions
3) Standard scores and the Normal Distribution
4) Point and interval estimation
5) Logic of single sample inference

The Normal Distribution

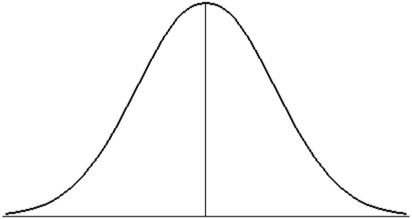
The normal distribution is characterized by two parameters: μ and σ

1) μ (mu) determines the **location** of a distribution

- Represents the population mean

2) σ (sigma) determines the **shape** of a distribution

- Represents the population standard deviation



Characteristics of the Normal Distribution

Three important features:

1) Symmetrical

- The vertical line through the center of the distribution (at μ) splits it into equal halves

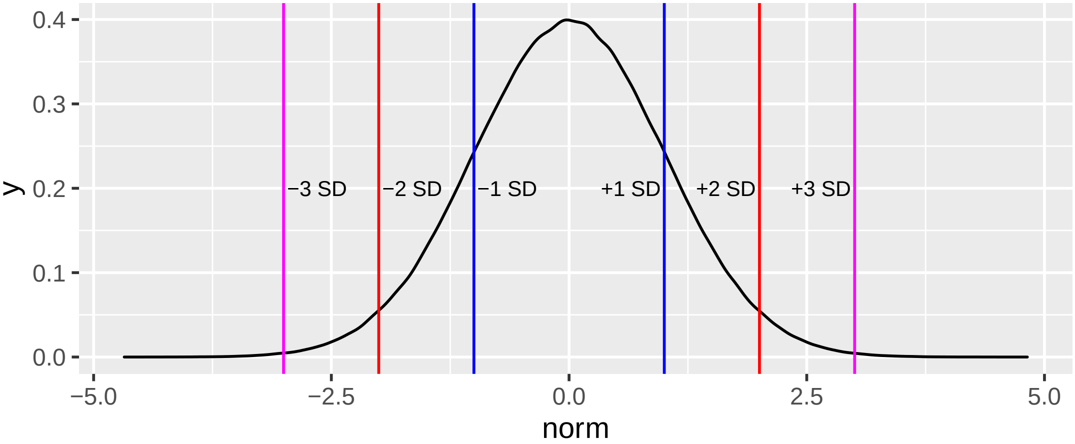
2) Unimodal

- The distribution has only one peak

3) Constant area

- No matter the shape of the distribution (i.e., regardless of σ) there is a constant amount of probability under the curve between the mean and any given distance from the mean measured in standard deviation units.

Characteristics of the Normal Distribution



Characteristics of the Normal Distribution

- No two normal distributions are exactly alike
  - Different μ and σ
  - Impossible to compare individual scores within and across distributions
- Solution is to transform the score by **standardizing** it.
- Transformed scores are called **z-scores** and are computed using the following equation:

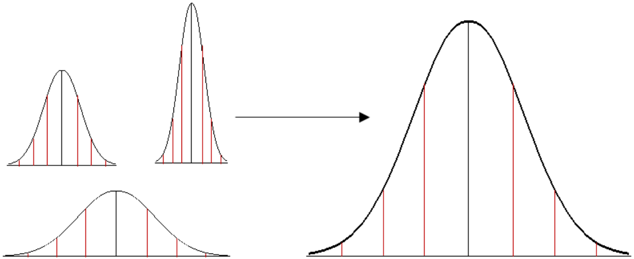
$$z = \frac{x - \mu}{\sigma}$$

  - z-scores tell us how many standard deviation units a particular observation lies from the mean

The Standard Normal Distribution

The Standard Normal Distribution contains a universal metric (hence the standard in its name).

This means that distributions of various sizes **can** now be compared, since they all share the same mean ($\mu = 0$) and standard deviation ($\sigma = 1$).



The Standard Normal Distribution

With this universal metric, we can now compute probabilities and assume them to be constant across all standard normal distributions.

z-Score	Area between μ and z	Area beyond z
0.00	.0000	.5000
0.50	.1915	.3085
1.00	.3413	.1587
1.50	.4332	.0668
2.00	.4772	.0228
2.50	.4938	.0062
3.00	.4987	.0013

The Standard Normal Distribution

Another representation of the z-table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.10	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.20	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.30	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.40	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319

Speeders Example

Let's say that the highway patrol collects data on cars traveling past a certain checkpoint.

The speed of cars traveling past this checkpoint is normally distributed with $\mu = 60.3$ and $\sigma = 6.5$.

Using these values, we can say something about the probability of observing different speeds (or ranges of speeds) if this distribution is representative of what we could expect among the population (for now we will assume it is for the sake of simplicity).

Speeders Example - Computing Standard Scores

What is the probability of observing a car going **more** than 70 miles per hour?

$$p(x > 70) \Rightarrow \frac{x-\mu}{\sigma} = \frac{70-60.3}{6.5} = 1.492 \Rightarrow p(z > 1.492) = 0.068$$

What is the probability of observing a car going **less** than 53 miles per hour?

$$p(x < 53) \Rightarrow \frac{x-\mu}{\sigma} = \frac{53-60.3}{6.5} = -1.123 \Rightarrow p(z < -1.123) = 0.131$$

Speeders Example - Computing Standard Scores

What is the probability of observing a car going **between** 61 and 63 miles per hour?

$$\mathrm{x} = 61 => \frac{61 - 60.3}{6.5} = 0.108; \;\; \mathrm{x} = 63 => \frac{63 - 60.3}{6.5} = 0.415$$

$$\Rightarrow p(61 < \mathrm{x} < 63) = p(0.108 < z < 0.415) = p(z > 0.108) - p(z > 0.415) = 0.118$$

Speeders Example - Computing Standard Scores

What is the probability of observing a car going **between** 60 and 65 miles per hour?

$$\text{x} = 60 => \frac{60 - 60.3}{6.5} = -0.046;\ \ \text{x} = 63 => \frac{65 - 60.3}{6.5} = 0.723$$

$$\Rightarrow p(60 < \text{x} < 65) = p(-0.046 < z < 0.723) = p(z > -0.046) - p(z > 0.723) = 0.283$$

Last answer will be slightly off (0.001) due to rounding error - using the qnorm function (demonstrated a bit further down below) will fix that.

A Note on the pnorm() Function

In the above calculations I use an R function called **pnorm**. This function returns the lower or upper-tail probability of finding some particular z-score, given a particular mean and standard deviation for a distribution. Instead of relying on inexact tables, it's more accurate to calculate a probability value in R using this function.

A Note on the pnorm() Function

The function takes the following general form:

pnorm(x, mean=0, sd=1, lower.tail=TRUE/FALSE)

Where x is the z-score you have obtained, mean=0 and sd=1 specify a standard normal distribution, and lower.tail=TRUE/FALSE tells R to return the lower (TRUE) or upper (FALSE) tail probability. Lower tail probabilities are to the LEFT of your z-score and upper tail probabilities are to the RIGHT of your z-score.

Speeders Example - Computing Standard Scores

We can also use algebra to work backwards from particular probabilities to obtain raw scores using the following equation:

$$x = (z * \sigma) + \mu$$

Suppose the county wants to aggressively enforce speed limits against the top 10% of speeders. Which speeds should they target to do so?

$$p(z > 1.282) = 0.1 => 1.282 = \frac{x - 60.3}{6.5} => x = (1.282 * 6.5) + 60.3 = 68.633$$

Additional Examples - Computing Standard Scores

In order to graduate with honors, students must be in the top 2% (summa cum laude), 3% (magna cum laude), or 5% (cum laude) of their graduating class.

Suppose that GPAs are normally distributed with mean $\mu = 2.60$ and standard deviation $\sigma = .65$. What GPA must a student have to graduate with each of these three honors?

Additional Examples - Computing Standard Scores

Summa Cum Laude:

$$p(z > 2.054) = 0.02 => 2.054 = \frac{\mathrm{x} - 2.60}{.65} => \mathrm{x} = (2.054 * 0.65) + 2.60 = 3.935$$

Magna Cum Laude:

$$p(z > 1.881) = 0.03 => 1.881 = \frac{\mathrm{x} - 2.60}{.65} => \mathrm{x} = (1.881 * 0.65) + 2.60 = 3.823$$

Cum Laude:

$$p(z > 1.645) = 0.05 => 1.645 = \frac{\mathrm{x} - 2.60}{.65} => \mathrm{x} = (1.645 * 0.65) + 2.60 = 3.669$$

A Note on the qnorm() Function

In the above calculations I use an R function called **qnorm**. This function returns a z-score that is associated with some lower/upper tail probability, given a particular mean and standard deviation for a distribution. Like the **pnorm** function, we can use the **qnorm** function to obtain more exact z-scores than we would if we were relying on z-score tables alone.

A Note on the qnorm() Function

The function takes the following general form:

qnorm(p, mean=0, sd=1, lower.tail=TRUE/FALSE)

Where p is the probability value you are interested in, mean=0 and sd=1 specify a standard normal distribution, and lower.tail=TRUE/FALSE tells R to return a z-score that leaves that value of probability to its left (TRUE) or right (FALSE). Because the normal distribution is symmetrical, this last part simply applies the right sign (postive/negative) to the z-score. You can test this out by holding all but the lower.tail option constant and shifting between lower.tail=TRUE and lower.tail=FALSE - the absolute value of the z-score remains the same, only its sign changes.

Logic of Sampling

Standard scores bring us one step closer toward being able to conduct hypothesis tests using a sample mean as an estimate for the population mean.

Our goal is to be able to use the sample mean as a **best guess** for the population mean.

Recall that one of the properties of a statistic, such as a sample mean, is that although it is empirical (i.e., it can be measured) and known (we actually collect data from a sample), it is not fixed.

Since we know this, we cannot say for certain that a sample statistic (e.g., $\overline{x}$) is equivalent to its corresponding population parameter (e.g., μ). Instead, we have to resort to what is known as a **sampling** distribution, which is just a particular type of probability distribution.

Logic of Sampling (cont.)

- Population mean (μ)
  - Empirical: can be measured
  - Unknown: Impractical to measure
  - Fixed: one *true* value (constant)
- Sample mean ($\overline{x}$)
  - Empirical: can be measured
  - Known: actually have data to measure it
  - Not fixed: varies from sample to sample (variable)

Sampling Error

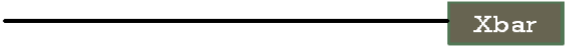
- Across samples, the estimate for μ will differ, even though samples are random draws from the **same population**.
- That is, sampling error produces randomness in the estimate of the mean
  - Within a population, the mean is a **constant**
  - Within a sample, the mean is a **constant**
  - Across samples, the mean is a **variable**

Sampling Error - Thought Experiment

Assume a random sample of size n, compute $\overline{x}$

Compute $\overline{x}$ from a second random sample, then third, fourth, fifth, etc. . . an infinite number of times.

What do you think is going to happen?



Sampling Distribution

- Distribution of all possible sample means
  - Theoretical (vs. empirical)
  - Known: probability theory
- Centered around the true population mean - μ

Properties of Sampling Distributions

- Our sample mean is but one mean from a theoretical distribution of all possible sample means
- We can then use the laws of probability to determine the probability of obtaining a *particular* sample mean
- Properties of sampling distributions:
  - As n increases, the standard error decreases
  - Sampling distribution is always normally distributed if the population values are normally distributed

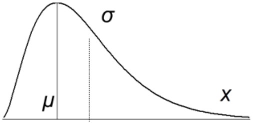
Central Limit Theorem

If an infinite number of sample sizes n are drawn from the population, the sampling distribution will approach normality as the sample size becomes infinitely large, **even if the characteristic is not normally distributed in the population**.

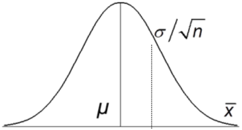
Even though a variable like *number of arrests* is highly skewed in the population, we can still assume that the sampling distribution will be normal when we have a large sample.

Central Limit Distribution (cont.)

- Population distribution (N observations)



- Sampling distribution (n observations)



Standard Score for a Sample Mean

To standardize a sample mean with respect to μ, we need a new calculation that standardizes it by **standard error** units:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

A New Example - Household Income

Assume the following population parameters: $\mu = \$32,000$ and $\sigma = \$5,000$. We take a sample of $n = 30$ and want to know the probability that we would observe $\overline{x}$ values such as:

$p(\overline{x} < \$30,000)$

A New Example - Household Income

Assume the following population parameters: $\mu = \$32,000$ and $\sigma = \$5,000$. We take a sample of $n = 30$ and want to know the probability that we would observe $\overline{x}$ values such as:

$p(\overline{x} < \$30,000)$

$$\begin{aligned} p(\overline{\mathrm{x}} < 30000) &= p(z < \frac{30000 - 32000}{5000/\sqrt{30}}) \\ &= p(z < -2.19) \\ &= 0.0143 \end{aligned}$$

A New Example - Household Income

Assume the following population parameters: $\mu = \$32,000$ and $\sigma = \$5,000$. We take a sample of $n = 30$ and want to know the probability that would would observe $\overline{x}$ values such as:

$p(\overline{x} > \$33,000)$

A New Example - Household Income

Assume the following population parameters: $\mu = \$32,000$ and $\sigma = \$5,000$. We take a sample of $n = 30$ and want to know the probability that would would observe $\overline{x}$ values such as:

$p(\overline{x} > \$33,000)$

$$p(\overline{\mathrm{x}} > 33000) = p(z > \frac{33000 - 32000}{5000/\sqrt{30}})$$

$$= p(z > 1.10)$$

$$= 0.1357$$

A New Example - Household Income

Assume the following population parameters: $\mu = \$32,000$ and $\sigma = \$5,000$. We take a sample of $n = 30$ and want to know the probability that would would observe $\overline{x}$ values such as:

$p(\$31,500 < \overline{x} < \$32,500)$

A New Example - Household Income

Assume the following population parameters: $\mu = \$32,000$ and $\sigma = \$5,000$. We take a sample of $n = 30$ and want to know the probability that would would observe $\overline{x}$ values such as:

$p(\$31,500 < \overline{x} < \$32,500)$

$$
\begin{aligned}
p(31500 < \overline{\mathrm{x}} < 32500) &= p(\frac{31500 - 32000}{5000/\sqrt{30}} < z < \frac{32500 - 32000}{5000/\sqrt{30}}) \\
&= p(-0.55 < z < 0.55) \\
&= 0.7082 - 0.2918 \\
&= 0.4163
\end{aligned}
$$

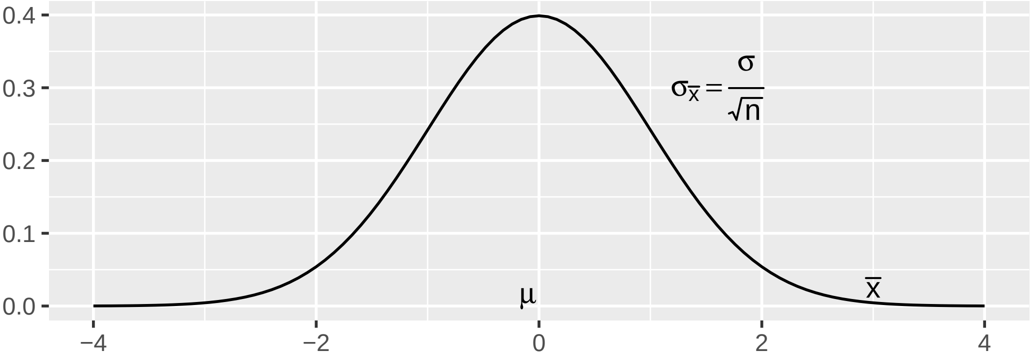
Sampling Distribution Redux

Just as a means of review (or to reinforce the above), here's a table listing the properties of the distributions we have talked about today:

Distribution	Properties	Mean	Standard Deviation
Sample	Empirical, known	$\overline{x}$	s
Population	Empirical, unknown	μ	σ
Sampling	Theoretical, known	μ	$\sigma_{\overline{x}} = \sigma / \sqrt{n}$

Central Limit Theorem Redux

With a large enough n, the sampling distribution is approximately normal, even if the characteristic is not normally distributed in the population.



Moving on - Point Estimation

What is a point estimate?

A point estimate is the **sample statistic** we use as an estimate of an *unknown* **population parameter**.

So, for example:

1) $\overline{x}$ is a point estimate for μ
2) s (the sample standard deviation) is a point estimate for σ

Properties of “Good” Point Estimates

We prefer that our point estimates have two properties: they are **unbiased** and they are **efficient**.

What does this mean, though?

Point Estimates - Unbiasedness

In order to be called **unbiased**, a point estimate must be equal to the true parameter being estimated.

$$E(\overline{x}) = E(x) = \mu$$

This does not mean that we will get the *true* answer every time, only that we will get it **on average**.

Point Estimates - Efficiency

In order to be called **efficient** the sampling distribution of point estimates should cluster tightly about the true population parameter.

This means that we will be close (i.e., minimal error) to the true answer, **on average**.

As n increases, $\overline{x}$ becomes a more **efficient** estimator for μ:

$$\sigma_{\overline{x}} = \sigma / \sqrt{n}$$

Interval Estimation

Point estimation leads us to internal estimation. We generally refer to these intervals as **confidence intervals**.

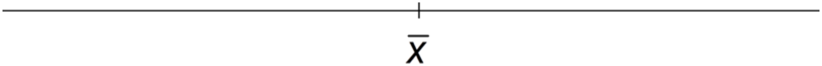
A **confidence interval** is a range of values within which a population parameter has a known probability of lying.

Interval Estimation

Since we know any **point estimate** contains **sampling error**, we take this uncertainty into account in the construction of the **confidence interval** when we estimate a range of values where the true population value *might* be.

Conventional levels of confidence map neatly onto typical values of statistical significance: 90% ($\alpha = .10$), 95% ($\alpha = .05$), 99% ($\alpha = .01$), and 99.9% ($\alpha = .001$).

Logic of Interval Estimation



Interpreting Confidence Intervals

- Repeated sampling interpretation of 95% CI
  - If we drew an infinite number of samples of size n from the population and constructed a 95% CI around each sample mean, 95% of these intervals would contain the true population mean
- Shorthand interpretation
  - We are 95% confident that the true population mean lies within the CI

Constructing Confidence Intervals

For these first few examples, we will operate under the assumption that we *know* the true value of σ for our calculation of confidence intervals. Here's the equation:

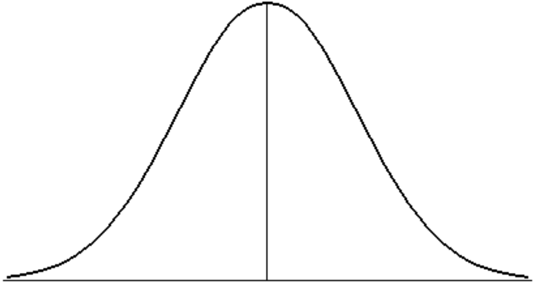
$$
\begin{aligned}
C.I. &= \overline{\mathrm{x}} \pm z_{\alpha/2}(\sigma_{\overline{\mathrm{x}}}) && (1)\\
&= \overline{\mathrm{x}} \pm z_{\alpha/2}(\sqrt{\sigma^2/\mathrm{n}}) && (2)\\
&= \overline{\mathrm{x}} \pm z_{\alpha/2}(\sigma/\sqrt{\mathrm{n}}) && (3)
\end{aligned}
$$

Note: $z_{\alpha/2}$ simply represents the z-score you would need for a two-tailed test at a particular α level, where the α level is equal to 1-confidence.

Finding "z" for a 95% CI

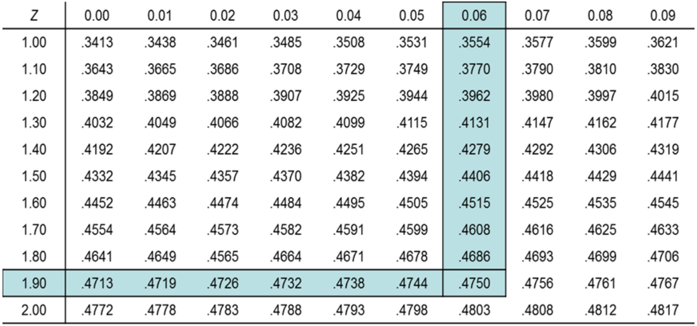
First, divide your confidence in half: $0.95/2 = .4750$.

This is the probability that lies between the mean and the z-score.



Finding “z” for a 95% CI (cont.)

Second, find the closest probability in the middle part of the standard normal (z) distribution.



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.00	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.10	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.20	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.30	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.40	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.50	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.60	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.70	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.80	.4641	.4649	.4565	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.90	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.00	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817

Finding “z” for a 95% CI

Third, identify the z-score = 1.96: 95% CI $= \overline{x} \pm 1.96(\sigma/\sqrt{n})$

In plain terms, this throws a *net* out 1.96 standard-error units from the sample mean.

We have 95% confidence that this *net* has captured the true population parameter being estimated (μ).

Let’s practice other confidence levels. . .

Finding “z” for a 95% CI (cont.)

- 90% CI
  - z =

Finding “z” for a 95% CI (cont.)

- 90% CI
  - z = 1.65
- 99% CI
  - z =

Finding “z” for a 95% CI (cont.)

- 90% CI
  - z = 1.65
- 99% CI
  - z = 2.58
- 99.9% CI
  - z =

Finding “z” for a 95% CI (cont.)

- 90% CI
  - $z = 1.65$
- 99% CI
  - $z = 2.58$
- 99.9% CI
  - $z = 3.27$

Conventional Confidence Levels

Confidence Level	Proportion from the Mean	Proportion in Tail	z-Score
90%	.4500	.0500	1.65
95%	.4750	.0250	1.96
99%	.4950	.0050	2.58
99.9%	.4995	.0005	3.27

An Alternative Method to find z

It is arguably much easier to simply use the qnorm function in R.

The important part to remember is that you will need to specify the tail proportion of the probability when using the qnorm function.

An Alternative Method to find z

So, to find the z-scores for a 95% confidence interval, I would use the following code:

```
qnorm(.0250, mean=0, sd=1, lower.tail=FALSE)
```

Question - why don't I have to figure out the z-score for the lower (left) tail?

Note: This will return a z-Score that is slightly below 1.96 - this is perfectly fine as it is more accurate than I can be in the above table.

CI Examples - Minneapolis Crime Hot Spots

- Calls to police for all addresses and intersections (*places*) in 1986
  - $n = 3795$ high-risk places
  - $\overline{x} = 43.03$ ($\sigma = 2.31$)
- 90% CI =

CI Examples - Minneapolis Crime Hot Spots

- Calls to police for all addresses and intersections (*places*) in 1986
  - $n = 3795$ high-risk places
  - $\overline{x} = 43.03$ ($\sigma = 2.31$)
- 90% CI =
  - $43.03 \pm 1.65(2.31/\sqrt{3795}) = 43.03 \pm .06 \Rightarrow [42.97, 43.09]$
- μ for Minnesota is actually 2.82

CI Examples - Minneapolis Crime Hot Spots

Let's increase our level of confidence and see what happens to the interval.

- 95% CI =

CI Examples - Minneapolis Crime Hot Spots

Let's increase our level of confidence and see what happens to the interval.

- 95% CI =
  - $43.03 \pm 1.96(2.31/\sqrt{3795}) = 43.03 \pm .07 \Rightarrow [42.96, 43.10]$
- 99% CI =

CI Examples - Minneapolis Crime Hot Spots

Let's increase our level of confidence and see what happens to the interval.

- 95% CI =
  - $43.03 \pm 1.96(2.31/\sqrt{3795}) = 43.03 \pm .07 \Rightarrow [42.96, 43.10]$
- 99% CI =
  - $43.03 \pm 2.58(2.31/\sqrt{3795}) = 43.03 \pm .10 \Rightarrow [42.93, 43.13]$
- 99.9% CI =

CI Examples - Minneapolis Crime Hot Spots

Let's increase our level of confidence and see what happens to the interval.

- 95% CI =
  - $43.03 \pm 1.96(2.31/\sqrt{3795}) = 43.03 \pm .07 \Rightarrow [42.96, 43.10]$
- 99% CI =
  - $43.03 \pm 2.58(2.31/\sqrt{3795}) = 43.03 \pm .10 \Rightarrow [42.93, 43.13]$
- 99.9% CI =
  - $43.03 \pm 3.27(2.31/\sqrt{3795}) = 43.03 \pm .12 \Rightarrow [42.91, 43.15]$

What do you think it means that our confidence intervals do not contain μ (2.82)?

CI Examples - US Homicide Rates

- Homicide rates per 100,000 since 1950
  - $n = 50$ states in 2017
  - $\overline{x} = 5.3$ $(\sigma = 1.97)$
- 95% CI =

CI Examples - US Homicide Rates

- Homicide rates per 100,000 since 1950
  - $n = 50$ states in 2017
  - $\overline{x} = 5.3$ $(\sigma = 1.97)$
- 95% CI =
  - $5.3 \pm 1.96(1.97/\sqrt{50}) = 5.3 \pm 0.55 \Rightarrow [4.75, 5.85]$
- The population mean (μ) is actually 6.6
- What does it mean that the true population mean is not contained in this interval?

Constructing Confidence Intervals - σ Unknown

Now let's consider the more typical alternative where σ is unknown.

Here's the equation for a CI where we do not know the population standard error:

$$\begin{aligned} \text{C.I.} &= \overline{\mathrm{x}} \pm t^{n-1}_{\alpha/2}(s_{\overline{\mathrm{x}}}) && (4) \\ &= \overline{\mathrm{x}} \pm t^{n-1}_{\alpha/2}(\sqrt{s^2/n-1}) && (5) \\ &= \overline{\mathrm{x}} \pm t^{n-1}_{\alpha/2}(s/\sqrt{n-1}) && (6) \end{aligned}$$

New probability distribution: Student's t (AKA, t-distribution)

Properties of the t-Distribution

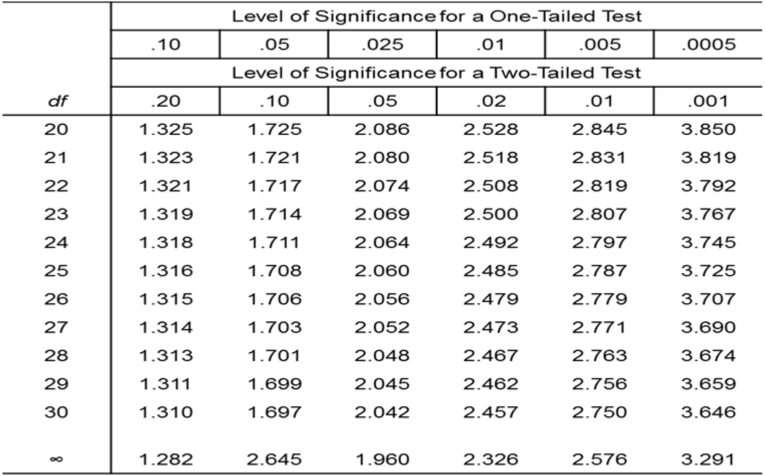
- Properties of a t-distribution
  - Approximately normal
  - Wider than the z-distribution
  - Fatter tails (i.e., more probability in the tails)
- Defined by **degrees of freedom** (df)
  - df = n-1
  - As n becomes infinitely large, the t-distribution converges to the z-distribution

Always use the t-distribution when σ is unknown (i.e., when we use s as a point estimate for σ).

t-Distribution (cont.)

- With small samples, t is larger than z
  - Reflects the fact that with small n there is greater sample to sample variability in the sample mean (i.e., more **sampling error**).
- Finding t
  - Determine the level of significance = 1-confidence (or just decide on the α level)
  - Always use two-tailed significance for a CI
  - Calculate degrees of freedom (*df*)
  - Identify t-score

Finding t



	Level of Significance for a One-Tailed Test					
	.10	.05	.025	.01	.005	.0005
	Level of Significance for a Two-Tailed Test					
df	.20	.10	.05	.02	.01	.001
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
∞	1.282	2.645	1.960	2.326	2.576	3.291

Finding t Using R

An alternative (and more accurate) way to find t-values is to use the qt() function. It's quite similar to the qnorm() function, but requires some additional information.

The default format of the function is:

qt(p, df, ncp, lower.tail=TRUE/FALSE, log.p=FALSE)

For our purposes, we will only use the p, df, and lower.tail options. The only new option is df, which stands for degrees of freedom.

Finding t Using R

As an example, let's say I want the t-value that leaves .05 probability points in either tail (so, a 90% CI) and I have a sample size of 30:

qt(.05, df=30-1, lower.tail=FALSE) = 1.6991

As a comparison, the corresponding function for a z-score looks like this:

qnorm(.05, mean=0, sd=1, lower.tail=FALSE) = 1.6449

Note - z will always be smaller than t - you can test this out by increasing the degrees of freedom and holding the probability value constant (try it for a sample size of 1000000 and see for yourself)

t-Distribution Examples - Prosocial Activities

- Prosocial activities among boot campers
  - $n = 1500$ youths
  - $\overline{x} = 1.53$ ($s = 1.21$)
- 99% CI =

t-Distribution Examples - Prosocial Activities

- Prosocial activities among boot campers
  - $n = 1500$ youths
  - $\overline{x} = 1.53$ ($s = 1.21$)
- 99% CI =
  - $1.53 \pm 2.576(1.21/\sqrt{1499}) = 1.53 \pm 0.08 \Rightarrow [1.45, 1.61]$

t-Distribution Examples - Prosocial Activities

- Prosocial activities among boot campers
  - $n = 1500$ youths
  - $\overline{x} = 1.53$ ($s = 1.21$)
- 99% CI =
  - $1.53 \pm 2.576(1.21/\sqrt{1499}) = 1.53 \pm 0.08 \Rightarrow [1.45, 1.61]$
- Now let's suppose that n=10 instead

t-Distribution Examples - Prosocial Activities

- Prosocial activities among boot campers
  - $n = 1500$ youths
  - $\overline{x} = 1.53$ $(s = 1.21)$
- 99% CI =
  - $1.53 \pm 2.576(1.21/\sqrt{1499}) = 1.53 \pm 0.08 \Rightarrow [1.45, 1.61]$
- Now let's suppose that n=10 instead
  - $1.53 \pm 3.25(1.21/\sqrt{9}) = 1.53 \pm 1.31 \Rightarrow [0.22, 2.84]$

t-Distribution Examples - Child Maltreatment

- Arrests among young adults with abuse history
  - $n = 61$
  - $\overline{x} = 2.57$ (s^2 =7.90)
- 95% CI =

t-Distribution Examples - Child Maltreatment

- Arrests among young adults with abuse history
  - $n = 61$
  - $\overline{x} = 2.57$ (s^2 =7.90)
- 95% CI =
  - $2.57 \pm 2(\sqrt{7.90/60}) = 2.57 \pm 0.73 \Rightarrow [1.84, 3.3]$

t-Distribution Examples - Child Maltreatment

- Arrests among young adults with abuse history
  - $n = 61$
  - $\overline{x} = 2.57$ (s^2 =7.90)
- 95% CI =
  - $2.57 \pm 2(\sqrt{7.90/60}) = 2.57 \pm 0.73 \Rightarrow [1.84, 3.3]$
- 99.9% CI =

t-Distribution Examples - Child Maltreatment

- Arrests among young adults with abuse history
  - $n = 61$
  - $\overline{x} = 2.57$ (s^2 =7.90)
- 95% CI =
  - $2.57 \pm 2(\sqrt{7.90/60}) = 2.57 \pm 0.73 \Rightarrow [1.84, 3.3]$
- 99.9% CI =
  - $2.57 \pm -3.46(\sqrt{7.90/60}) = 2.57 \pm 1.26 \Rightarrow [1.31, 3.83]$
- Notice that increased confidence widens the interval - increased confidence comes with the price of less precision.

t-Distribution Examples - Child Maltreatment

- Now, suppose $n = 250$
- 95% CI $=$

t-Distribution Examples - Child Maltreatment

- Now, suppose $n = 250$
- 95% CI =
  - $2.57 \pm 1.97(\sqrt{7.90/249}) = 2.57 \pm 0.351 \Rightarrow [2.219, 2.921]$

t-Distribution Examples - Child Maltreatment

- Now, suppose $n = 250$
- 95% CI $=$
  - $2.57 \pm 1.97(\sqrt{7.90/249}) = 2.57 \pm 0.351 \Rightarrow [2.219, 2.921]$
- 99.9% CI $=$

t-Distribution Examples - Child Maltreatment

- Now, suppose $n = 250$
- 95% CI =
  - $2.57 \pm 1.97(\sqrt{7.90/249}) = 2.57 \pm 0.351 \Rightarrow [2.219, 2.921]$
- 99.9% CI =
  - $2.57 \pm 3.33(\sqrt{7.90/249}) = 2.57 \pm 0.59 \Rightarrow [1.98, 3.16]$
- What do you observe about the size of these CIs compared to when the sample size was 61?
- That's right, the intervals decreased in width - remember that larger samples mean more efficient point estimation. This translates to thinner intervals, all else equal.

Confidence Intervals - Confidence v. Precision

- As a general rule, there is a trade-off between confidence and precision.
  - $\Uparrow$ confidence, $\Downarrow$ precision
  - $\Uparrow$ precision, $\Downarrow$ confidence
- No such thing as a free lunch!
  - Being more confident comes at a price: larger interval width.

Whew. . . Now for More

If interval estimation seems familiar to hypothesis testing, this is no coincidence.

Basically, interval estimation amounts to placing exact boundaries on the point estimates we *think* we should obtain within some distance from the mean.

If we happen to obtain **sample statistics** outside this realm, we would say that the result is **statistically significant**.

In other words, if the **population parameter** is true, we would expect to observe a **sample statistic** such as ours infrequently enough such that we are comfortable saying our sample may not come from the population defined by that parameter.

Logic of Single Sample Inference

More formally, interval estimation and hypothesis testing represent two sides of the very same coin.

In interval estimation, we use $\overline{x}$ to say something about μ.

More specifically, with interval estimation we use $\overline{x}$ as a **point estimate** in order to estimate a *plausible range* of values for μ.

Logic of Single Sample Inference

By contrast, with hypothesis testing we:

1) Make an assumption about what the true value of μ is (sometimes we know for sure).
2) Determine where, in the sampling distribution centered around μ does $\overline{x}$ fall (i.e., how many $\sigma_{\overline{x}}$ units is it away from μ).
3) Conclude whether our assumption about μ is realistic (in a probabilistic sense) given the statistic obtained from our sample.

Question of Single Sample Inference

Is it likely that our sample was drawn from a population with mean μ, or is our sample a subset of an entirely different population?

In other words, is μ the mean of the sampling distribution that this sample mean came from?

Two Possible Answers

- Yes, our sample was drawn from the population of interest.
  - The observed difference between $\overline{x}$ and μ is due to sampling error (i.e., a **chance** difference).
- No, our sample was drawn from a different population altogether.
  - Observed difference between $\overline{x}$ and μ is due to the fact that there are actually two different populations with two different means (i.e., a **systematic** difference).

Single Sample Inference

Inherently a probabilistic question - What is the probability we would obtain the observed sample mean if μ was actually true? I.e., is this difference random or systematic?

We need to use the calculations (z/t-scores) in this lecture to transform the sample statistics from varied samples (i.e., $\overline{x}$ and s) into a common metric.

Single Sample Inference Example - Minneapolis Crime Hot Spots

- Calls to police for all addresses and intersections (*places*) in 1986
  - $\mu = 2.82$ ($\sigma = 2.31$)
  - $n = 3795$ high-risk places
  - $\overline{x} = 43.03$
- Research Question
  - Do high-risk areas make more calls to the police than the general population?

Single Sample Inference - Hypothesis Test Steps

Step 1: Formally State Hypotheses

1) Alternative (research) hypothesis (H_1)
   - Answers research question in the affirmative
   - Can be directional (one-tailed) or non-directional (two-tailed)
2) Null hypothesis (H_0)
   - Answers research question in the negative

$$H_1 : \mu > 2.82$$

$$H_0 : \mu \leq 2.82$$

Single Sample Inference - Hypothesis Test Steps

Step 2: Obtain a probability distribution

Is σ known or unknown?

If **known**, we choose the **z-distribution**.

If **unknown**, we choose the **t-distribution**.

Single Sample Inference - Hypothesis Test Steps

Step 3: Make decision rules

Level of significance (α): one v. two-tailed

- Critical values rejection region
  - How many standard error units away?
- Current test:
  - $\alpha = .01$ (one-tailed)
  - $z_{crit} = 2.33$
  - Reject H_0 if TS > 2.33

Single Sample Inference - Hypothesis Test Steps

Step 4: Compute test statistic

Standard score for the mean (σ known):

$$TS = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{43.03 - 2.82}{2.31/\sqrt{3795}} = \frac{40.21}{.037} = 1086.76$$

What does this mean?

Single Sample Inference - Hypothesis Test Steps

Step 5: Make a decision about the null hypothesis

If TS is in critical region, reject H_0.

If TS is in acceptance region, accept H_0

Our decision: Reject H_0 since 1086.76>2.33

- Conclusion: $\mu > 2.82$
  - There are significantly more calls to the police in high-risk places.

Step 5 - Potential Errors

	Decision	
Reality	Accept H_0	Reject H_0
H_0 is true	Correct	Type I Error
H_0 is false	Type II Error	Correct

Type I and Type II Errors

Type I errors may be reduced by decreasing the level of significance (α).

- In this, these types of errors are inversely related:
  - $\Downarrow p(\text{Type I Error}), \Uparrow p(\text{Type II Error})$
  - $\Downarrow p(\text{Type II Error}), \Uparrow p(\text{Type I Error})$

Additional Example - US Homicide Rates

- Homicide rates per 100,000 since 1950
  - $\mu = 6.66$ ($\sigma = 1.97$)
  - $n = 50$ states in 2017
  - $\overline{x} = 5.33$
- Research Question
  - Has there been a change from the average homicide rate over the last ~70 years?

US Homicide Rates - Hypothesis Test

- **Step 1: Formally state hypotheses**

US Homicide Rates - Hypothesis Test

- **Step 1: Formally state hypotheses**
  - $\mathrm{H}_1 : \mu \neq 6.66; \mathrm{H}_0 : \mu = 6.66$
- **Step 2: Obtain a probability distribution**

US Homicide Rates - Hypothesis Test

- **Step 1: Formally state hypotheses**
  - $H_1 : \mu \neq 6.66$; $H_0 : \mu = 6.66$
- **Step 2: Obtain a probability distribution**
  - z-distribution
- **Step 3: Make decision rules**

US Homicide Rates - Hypothesis Test

- **Step 1: Formally state hypotheses**
  - $\mathrm{H}_1 : \mu \neq 6.66$; $\mathrm{H}_0 : \mu = 6.66$
- **Step 2: Obtain a probability distribution**
  - z-distribution
- **Step 3: Make decision rules**
  - $\alpha = .05$(two-tailed); $z_{\text{crit}} = 1.96$; reject H_0 if $|\mathrm{TS}| > 1.96$

US Homicide Rates - Hypothesis Test

- **Step 4: Calculate test statistic**

US Homicide Rates - Hypothesis Test

- **Step 4: Calculate test statistic**

$$\text{TS} = \frac{5.33 - 6.66}{1.97/\sqrt{50}} = \frac{-1.33}{.2786} = -4.77$$

- **Step 5: Make a decision about the null hypothesis**
  - Reject H_0, there was a significant change in the homicide rate in 2017.

Unknown Population σ

- When we do not know σ, we must use s as a point estimate for it.
  - Since we have to estimate this additional parameter, we take on additional sampling error.
  - That is, there is now sampling error in our point estimate for **both** μ and σ
- We must now adjust for degrees of freedom (df)
  - $df = n - 1$

Unknown Population σ (cont.)

Further, when we do not know σ we must use an alternative calculation for the test statistic:

$$\mathrm{TS} = \frac{\bar{\mathrm{x}} - \mu}{\sqrt{s^2/n-1}} = \frac{\bar{\mathrm{x}} - \mu}{s/\sqrt{n-1}}$$

Probability distribution is now t instead of z

t-Distribution Example - Training School

- Research Question
  - Do training school youth offend more than non-training school youth?
- Sample details
  - $\mu = 0.75$
  - $n = 121$
  - $\overline{x} = 2.25$ $(s^2 = 12.51)$

t-Distribution Example - Training School

- **Step 1: Formally state hypotheses**
  - $\text{H}_1 : \mu > 0.75$; $\text{H}_0 : \mu \leq 0.75$
- **Step 2: Obtain probability distribution**
  - t-distribution
- **Step 3: Make decision rules**
  - $\alpha = .01$(one-tailed); $t_{\text{crit}} = 2.36$; reject H_0 if TS > 2.36

t-Distribution Example - Training School

- **Step 4: Calculate test statistic**
  - $\text{TS} = \frac{\bar{\text{x}} - \mu}{\sqrt{s^2/n-1}} = \frac{2.25-0.75}{\sqrt{12.51/(121-1)}} = \frac{1.5}{.323} = 4.64$
- **Step 5: Make a decision about the null hypothesis**
  - Reject H_0; training school youths commit significantly more offenses than the overall population of youths.

t-Distribution Example - Childhood Maltreatment

- Research Question
  - Is there an association between childhood maltreatment and arrest as an adult?
- Sample details
  - $\mu = 1.25$
  - $n = 61$
  - $\overline{x} = 2.57$, $s = 2.81$

t-Distribution Example - Childhood Maltreatment

- **Step 1: Formally state hypotheses**
  - $\mathrm{H}_1 : \mu \neq 1.25$; $\mathrm{H}_0 : \mu = 1.25$
- **Step 2: Obtain probability distribution**
  - t-distribution
- **Step 3: Make decision rules**
  - $\alpha = .05$(two-tailed); $t_{\text{crit}} = 2.0003$; reject H_0 if TS > 2.0003

t-Distribution Example - Childhood Maltreatment

- **Step 4: Calculate test statistic**
  - TS $= \frac{\overline{\mathrm{x}}-\mu}{\sqrt{s^2/n-1}} = \frac{2.57-1.25}{2.81/\sqrt{(61-1)}} = \frac{1.32}{.363} = 3.64$
- **Step 5: Make a decision about the null hypothesis**
  - Reject H_0; child abuse victims are arrested a significantly *different* number of times than individuals who were not abused as children.

On p-values

An additional piece of information we can obtain about a z or t-score is the probability value associated with it.

These probability values indicate the probability of observing a test statistic that or more extreme given the assumption about the null hypothesis is true.

On p-values

In the training school example, our TS was 4.64. We conducted a one-tailed significance test. This gives us a probability value of 0.000004.

In the childhood maltreatment example, the TS was 3.64. We conducted a two-tailed significance test so we must multiply our probability by two (because the critical regions are in opposite tails of the distribution). This gives us a probability value of 0.0006.

On p-values

What you may also notice is that, if we reject H_0 we will **always** have a probability value lower than our selected α level. If we fail to reject, it will be equal to or greater than α.

Two Questions

Time for your two questions!