STAT 113
Analytic Inference for a Single Proportion

Colin Reimer Dawson

Oberlin College

7-10 April 2017

Outline

Outline

Limits of Normal Approximation So Far

- We have still needed to do all that randomization / resampling to calculate the standard error.
- We can avoid that with some more theory.

Cases to Address

We will need standard errors to do CIs and tests for the following parameters:

1. Single Proportion (now)
2. Single Mean (today)
3. Difference of Proportions (Thursday)
4. Difference of Means (Thursday)
5. Mean of Differences (new! next week)

Analytic Approximations of Sampling Distributions

Param.	Stat.	Randomization	Theory SE	Test Dist.
p	$\hat{p}$	Simulate from p_0	$\sqrt{\frac{p_0(1-p_0)}{n}}$	Normal
μ	$\bar{x}$	Bootstrap + shift	$\frac{s}{\sqrt{n}}$	t_{n-1}
$p_A - p_B$	$\hat{p}_A - \hat{p}_B$	Scramble groups	$\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$	Normal
$\mu_A - \mu_B$	$\bar{x}_A - \bar{x}_B$	Scramble groups	$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$	$t_{\min(n_A-1, n_B-1}$
μ_D	$\bar{x}_D$	Flip pairs*	$\frac{s_D}{\sqrt{n_D}}$	t_{n_D-1}
ρ	r	Scramble pairings	$\sqrt{\frac{1-r^2}{n-2}}$	t_{n-2}

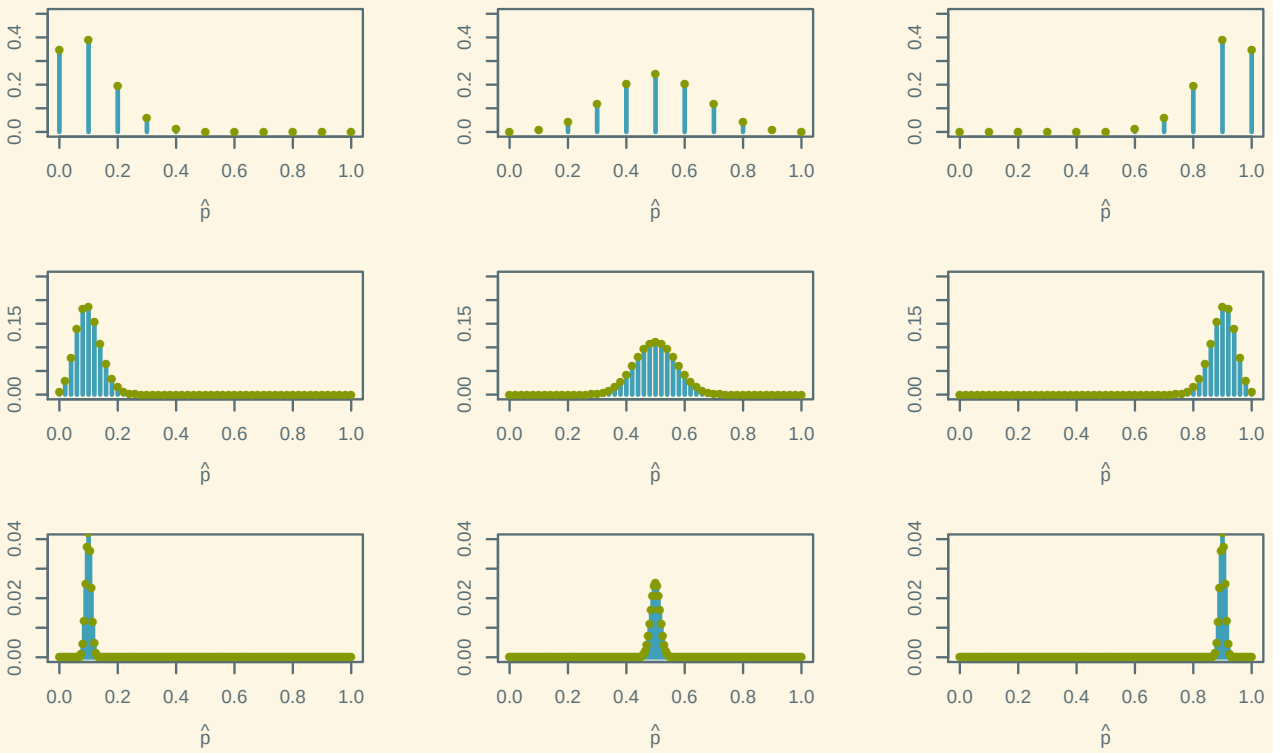
$$CI: \quad \text{Statistic} \pm \text{Critical Value} \times \widehat{SE}$$

$$\text{Sandardized Test Statistic}: \quad \frac{\text{Statistic} - \text{Null Param.}}{\widehat{SE}}$$

Outline

Outline

Sampling Distribution of a Sample Proportion



Columns: values of p (left: 0.1, middle: 0.5; right: 0.9)
Rows: values of n (top: 10, middle: 50; bottom: 1000)

Things Affecting the Standard Error for $\hat{p}$

1. Sample Size (n)
   - Increasing n makes the standard error go _____
2. Population Proportion (p)
   - What values of p make SE larger?

Distribution of $\hat{p}$

- Condition: The sampling distribution of $\hat{p}$ is approximately Normal with at least 10 expected cases of each outcome:

$$np \geq 10 \qquad n(1-p) \geq 10$$

- Mean: p
- Standard deviation (standard error):

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Outline

CI Summary: Single Proportion

To compute a confidence interval for a proportion when the bootstrap distribution for $\hat{p}$ is approximately Normal (i.e., counts for both outcomes ≥ 10), use

$$\hat{p} \pm Z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where Z^* is the Z-score of the endpoint appropriate for the confidence level, computed from a standard normal ($\mathcal{N}(0,1)$).

Example: Kissing Right

Most people are right-handed, and even the right eye is dominant for most people. Developmental biologists have suggested that late-stage human embryos tend to turn their heads to the right. In a study reported in Nature (2003), German bio-psychologist Onur Güntürkün studied kissing couples in public places such as airports, train stations, beaches, and parks. They observed 124 couples, age 13-70 years. For each kissing couple observed, the researchers noted whether the couple leaned their heads to the right or to the left.

Let's find a 95% confidence interval for p, the proportion of all couples who lean right.

Outline

P-values for a sample proportion from a Standard Normal

Computing P-values when the null sampling distribution is approximately Normal (i.e., np_0 and $np_0(1-p_0) \geq 10$) is the reverse process:

1. Convert $\hat{p}$ to a z-score within the theoretical distribution .

$$Z_{observed} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

2. Find the relevant area beyond $Z_{observed}$ using a Standard Normal

Example: Kissing Right

Most people are right-handed, and even the right eye is dominant for most people. Developmental biologists have suggested that late-stage human embryos tend to turn their heads to the right. In a study reported in Nature (2003), German bio-psychologist Onur Güntürkün studied kissing couples in public places such as airports, train stations, beaches, and parks. They observed 124 couples, age 13-70 years. For each kissing couple observed, the researchers noted whether the couple leaned their heads to the right or to the left.

Let's assess how strong the evidence is against the null hypothesis that couples are equally likely to lean right and left.

Outline

Outline

Distribution of Sample Means

- Central Limit Theorem: Sampling Distribution of $\bar{x}$ is approximately Normal, for "sufficiently large" samples, or when the population distribution is Normal.
- As the sample size n goes up, the standard error goes _____.
- Pairs: What effect do you expect the *population standard deviation* to have on the standard error of the distribution of sample means? Why?

Distribution of $\bar{x}$

- Population with mean μ and standard deviation σ
- Conditions: Sampling distribution of $\bar{x}$ is Normal if
  - Population is Normal, or
  - Sample size is large (roughly can use $n \geq 27$)
- Mean: μ
- Standard deviation (standard error):

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Outline

CI Summary: Single Mean

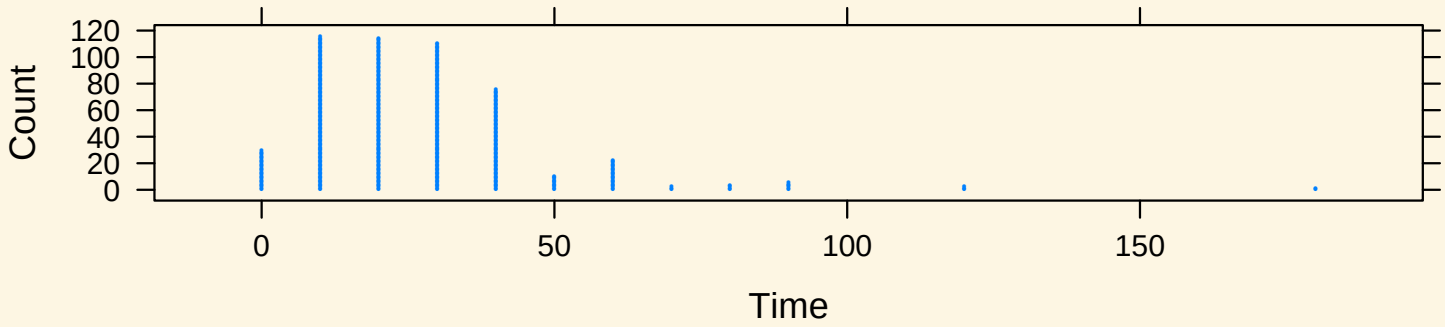
To compute a confidence interval for a mean when the sampling distribution for $\bar{x}$ is approximately Normal (i.e., Normal population, or "large" n), use

$$\bar{x} \pm Z^* \cdot \frac{\sigma}{\sqrt{n}}$$

where Z^* is the Z-score of the endpoint appropriate for the confidence level, computed from a standard normal ($\mathcal{N}(0, 1)$).

Example: Mean Atlanta Commute Time

```
library("mosaic"); library("Lock5Data"); data("CommuteAtlanta")
dotPlot(~Time, data = CommuteAtlanta, width = 10, cex = 4)
```




```
nrow(CommuteAtlanta)
```

```
[1] 500
```

```
mean(~Time, data = CommuteAtlanta)
```

```
[1] 29.11
```

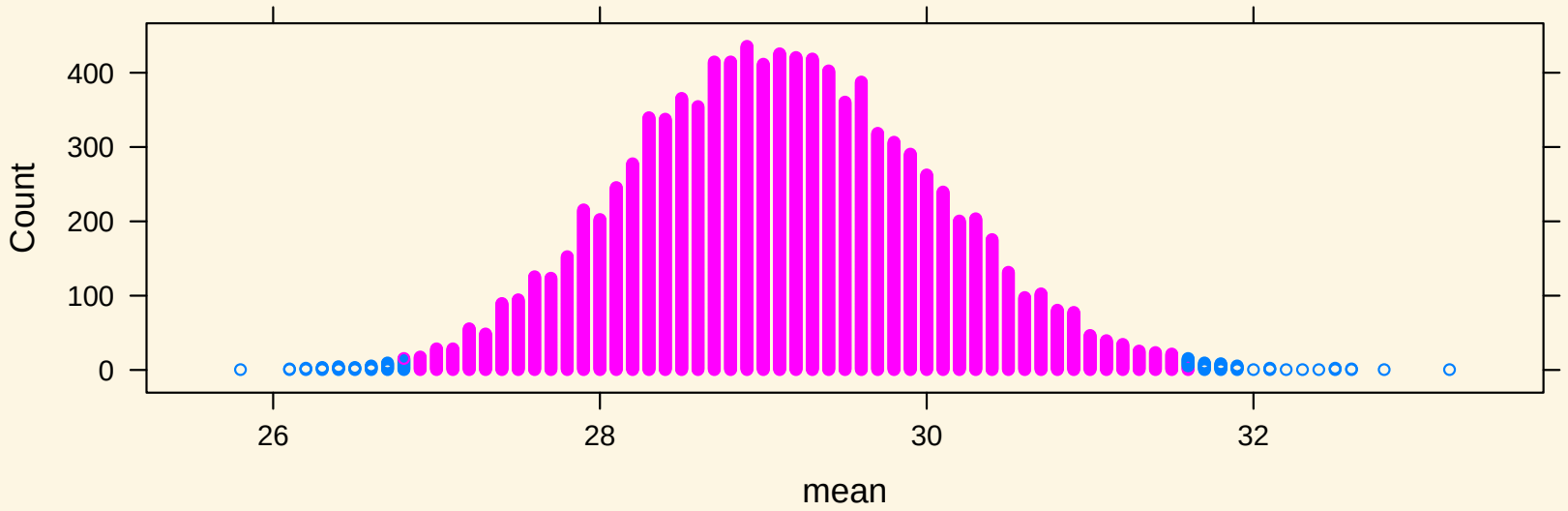
Atlanta Commute Time: Bootstrap CI

```
Bootstrap.means <- do(10000) * mean(~Time, data = resample(CommuteAtlanta))
CI.99.boot <-
    quantile(~mean, data = Bootstrap.means, prob = c(0.005, 0.995))
CI.99.boot

##     0.5%    99.5%
## 26.84399 31.58002
```

Commute Time: Pure Bootstrap CI

```
dotPlot(~mean, data = Bootstrap.means, width = 0.1, cex = 20,
        groups = mean >= CI.99.boot[1] & mean <= CI.99.boot[2])
```



Atlanta Commute Time: Analytic CI

- Confidence interval

$$\bar{x} \pm Z^* \cdot SE$$

- $\bar{x} = 29.11$
- $Z^* \approx 1.96$
- SE: $\frac{\sigma}{\sqrt{n}}$
- $n = 500$
- Wait, where do we get σ?

Outline

Using s instead of σ

- We can approximate SE with $\frac{s}{\sqrt{n}}$, but need to account for the fact that s itself is an estimate (differing between samples).
- "95% of sample means are within 2SE of μ" no longer accurate: the percentage is less than this.
- How much less depends on how good an estimate s is of σ (i.e., depends on n).

Degrees of Freedom

Recall

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$n-1$ is the "degrees of freedom", or the number of "pieces of information" we have about variability.

Bigger $df \rightarrow$ more accurate reflection of σ.

The t family of distributions

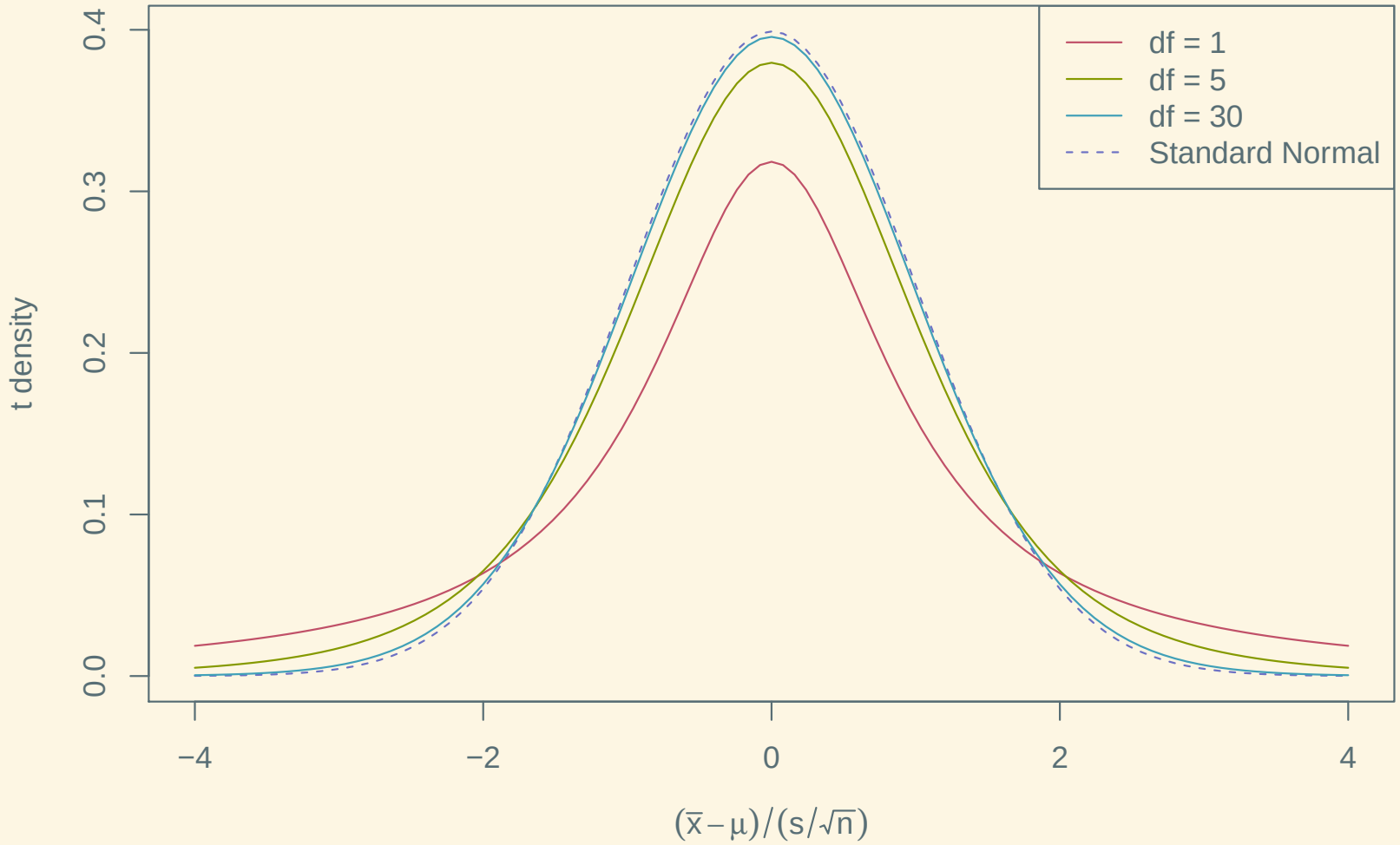
When we know σ, we have

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

i.e., z-scores calculated from sample means have a Standard Normal
When we don't know σ (almost always), estimate with s, then

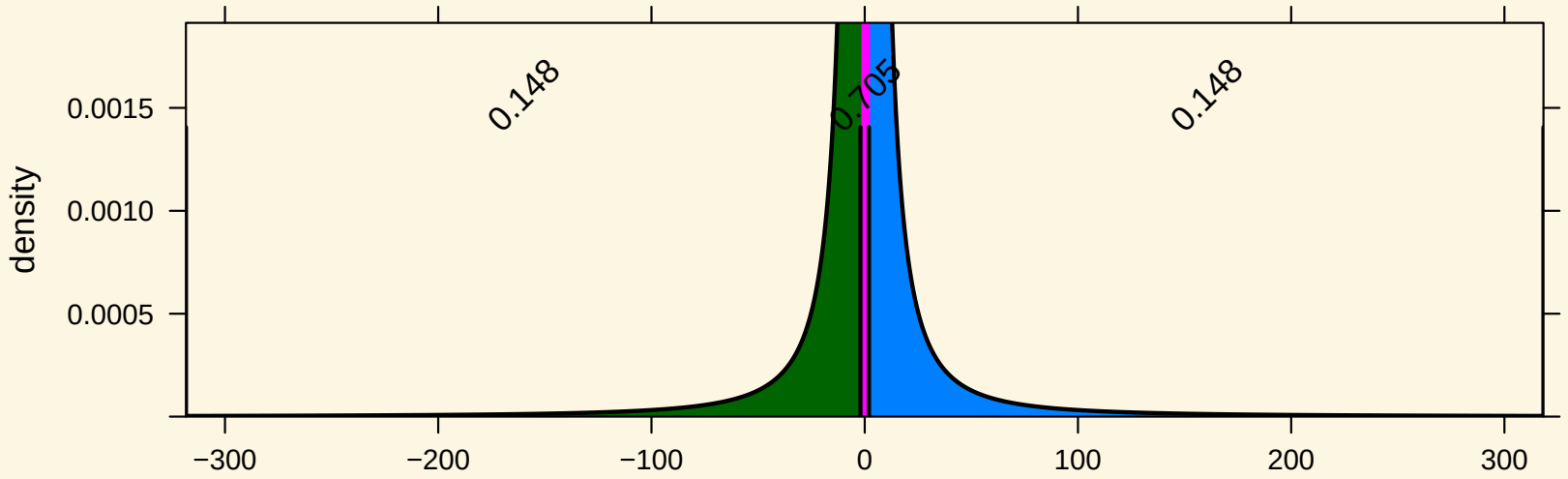
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

A family of t distributions



Tail Probabilities in t distributions

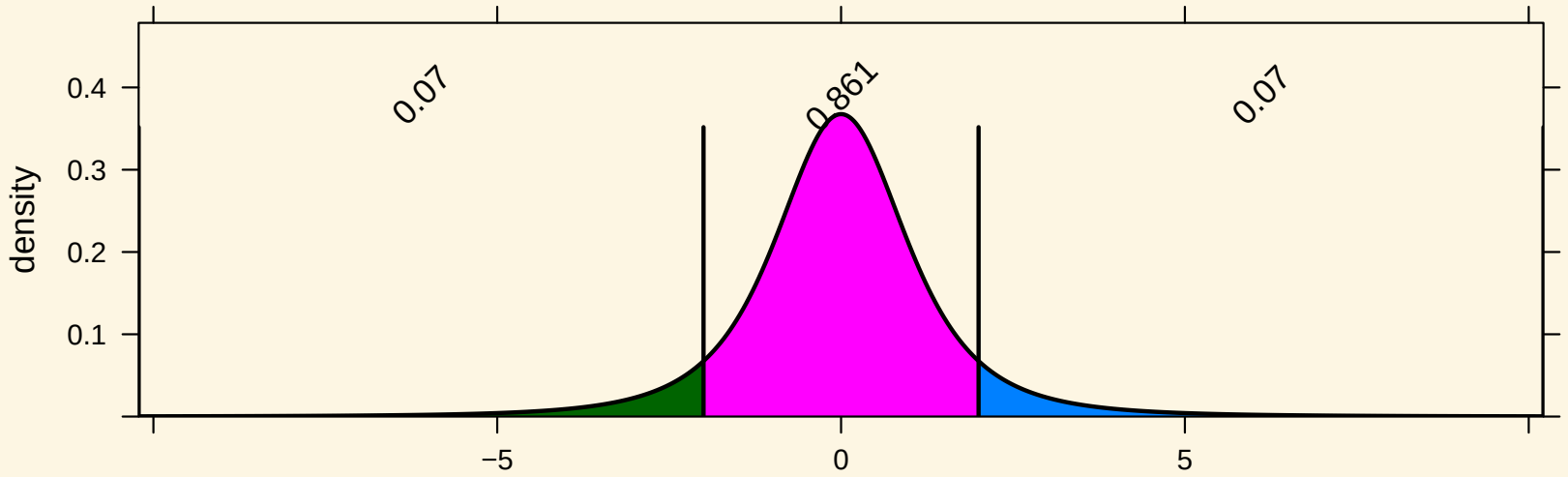
```
xpt(c(-2, 2), df = 1)
```




```
[1] 0.1475836 0.8524164
```

Tail Probabilities in t distributions

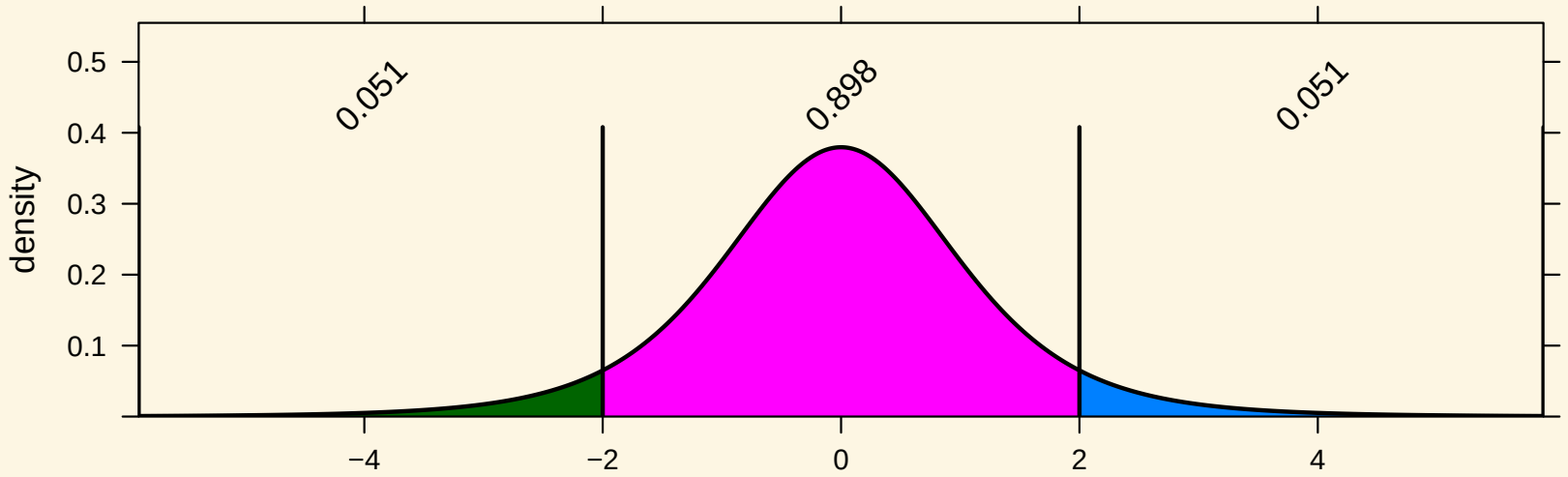
```
xpt(c(-2, 2), df = 3)
```




```
[1] 0.06966298 0.93033702
```

Tail Probabilities in t distributions

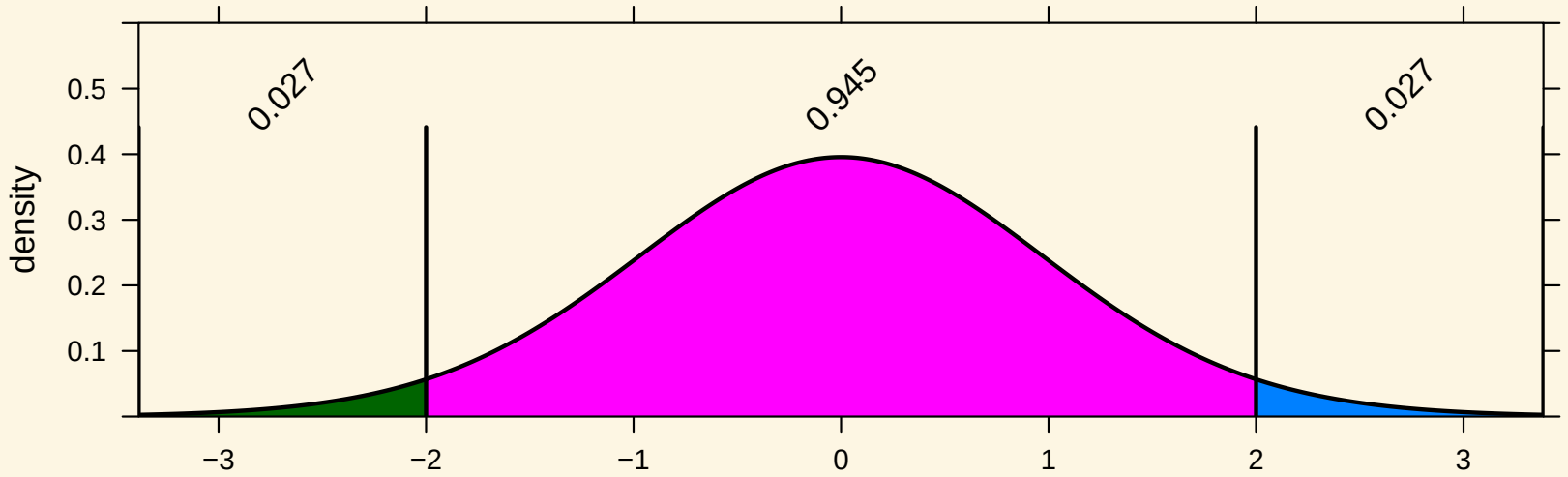
```
xpt(c(-2, 2), df = 5)
```




```
[1] 0.05096974 0.94903026
```

Tail Probabilities in t distributions

```
xpt(c(-2, 2), df = 30)
```



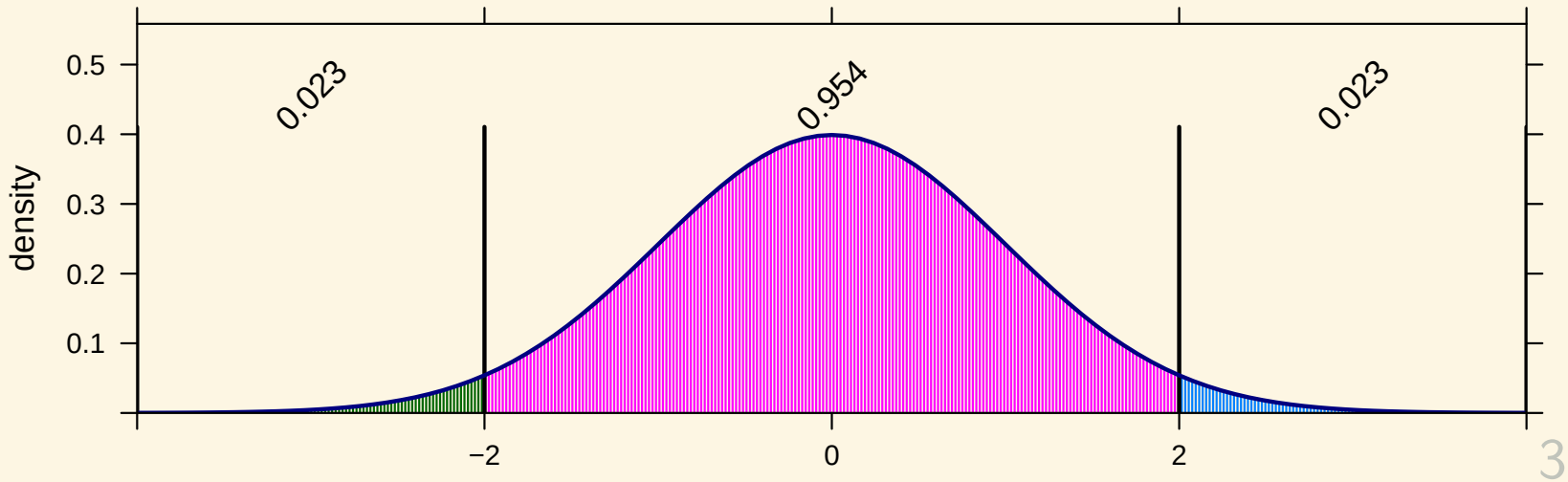

```
[1] 0.02731252 0.97268748
```

Tail Probabilities in Standard Normal distribution

```
xpnorm(c(-2, 2))
```

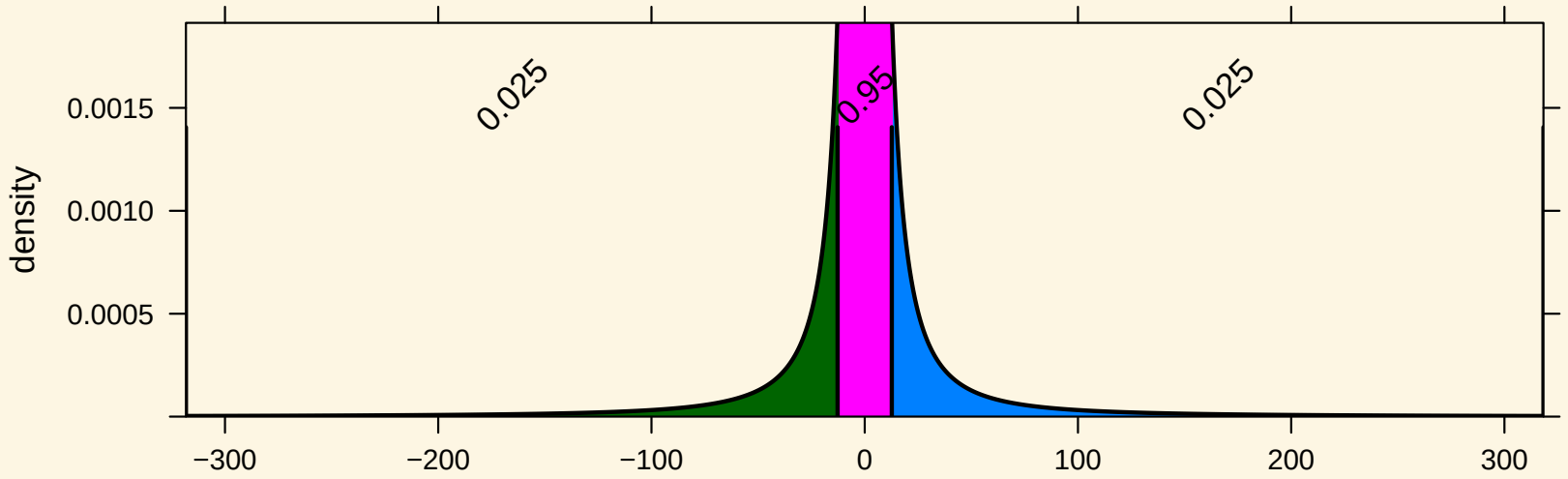
```
     If X ~ N(0, 1), then

      P(X <= -2) = P(Z <= -2) = 0.02275013
       P(X <=  2) = P(Z <=  2) = 0.97724987
      P(X >  -2) = P(Z >  -2) = 0.97724987
       P(X >   2) = P(Z >   2) = 0.02275013
```



Quantiles of t distributions

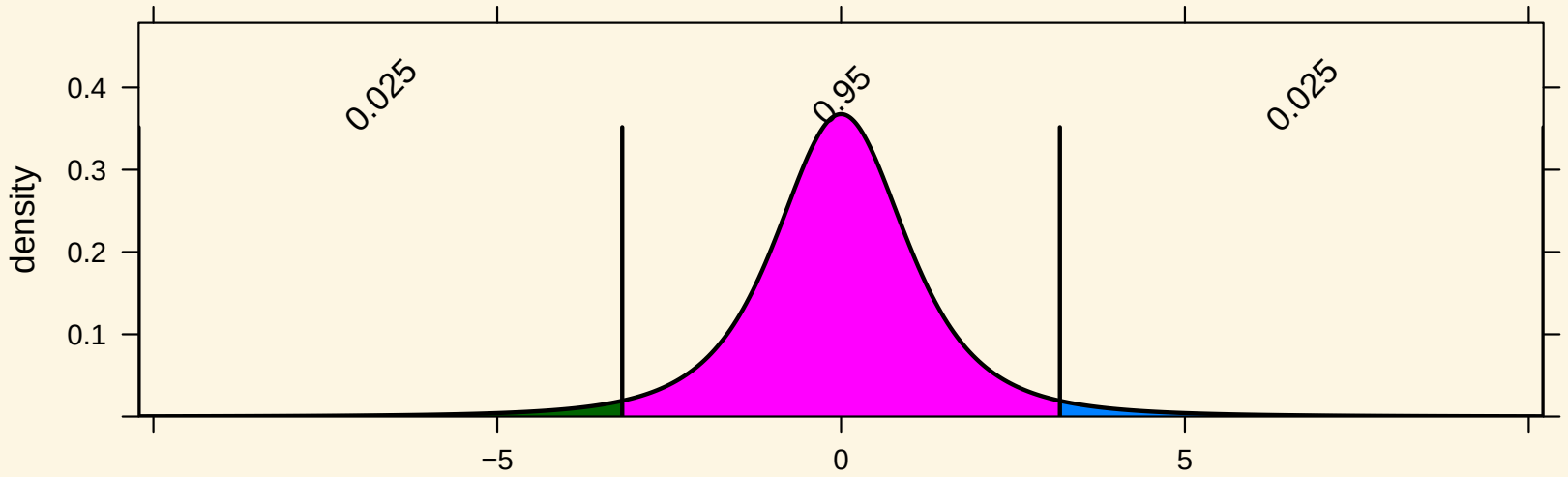
```
xqt(c(0.025, 0.975), df = 1)
```




```
[1] -12.7062  12.7062
```

Quantiles of t distributions

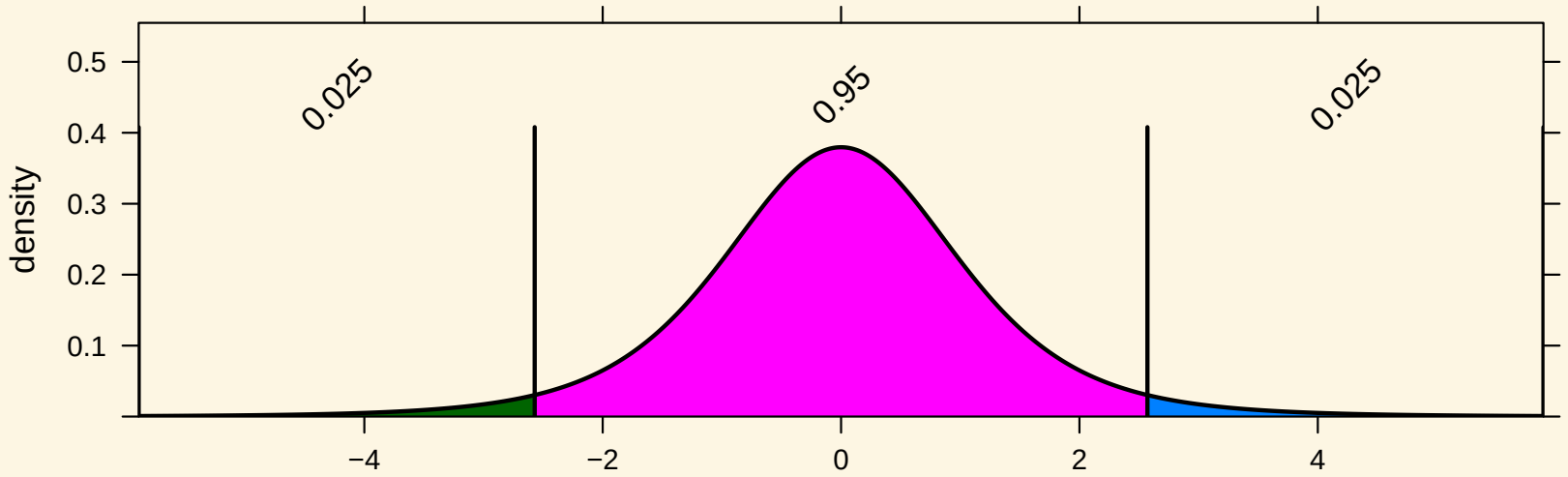
```
xqt(c(0.025, 0.975), df = 3)
```




```
[1] -3.182446  3.182446
```

Quantiles of t distributions

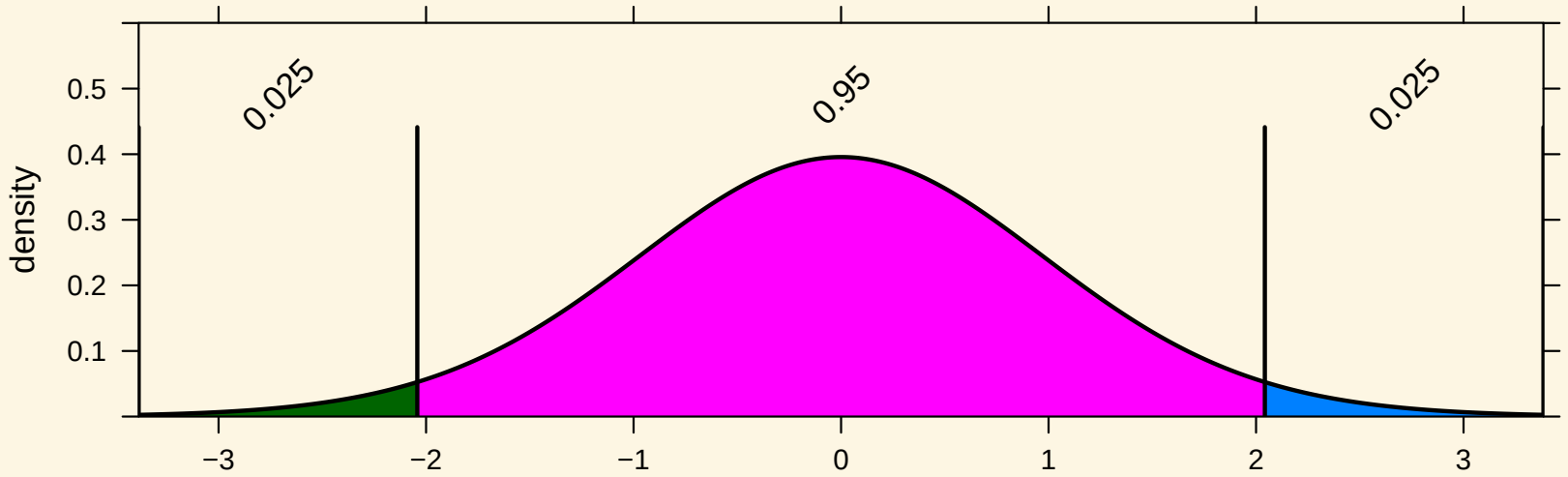
```
xqt(c(0.025, 0.975), df = 5)
```




```
[1] -2.570582  2.570582
```

Quantiles of t distributions

```
xqt(c(0.025, 0.975), df = 30)
```

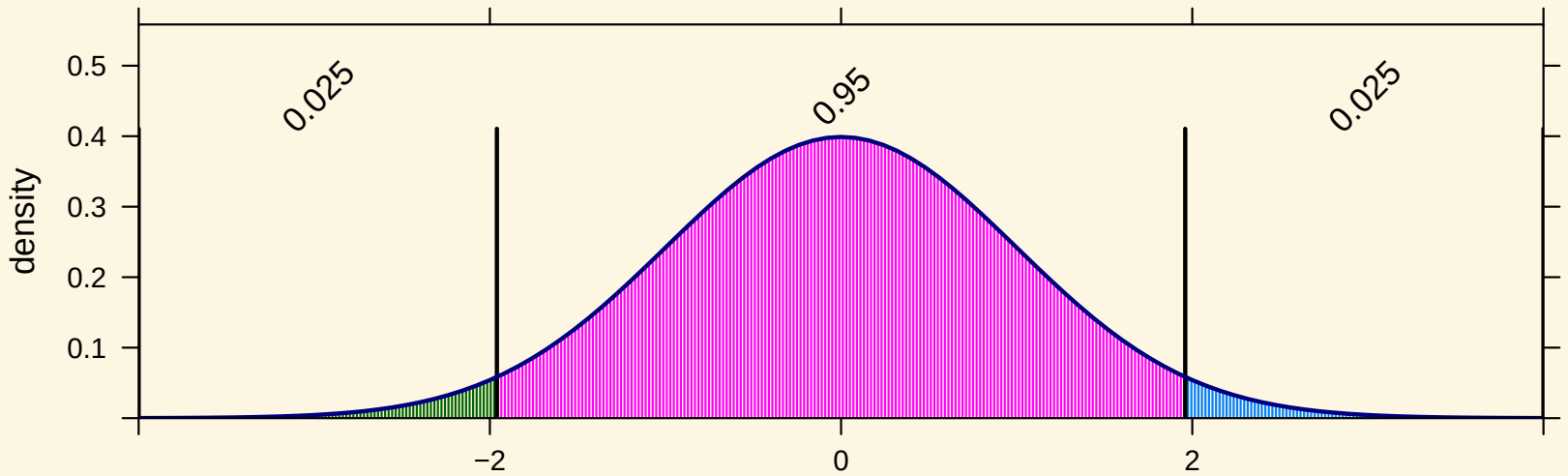



```
[1] -2.042272  2.042272
```

Quantiles of Standard Normal distribution

```
xqnorm(c(0.025, 0.975))

      P(X <= -1.95996398454005) = 0.025
       P(X <= 1.95996398454005) = 0.975
      P(X >  -1.95996398454005) = 0.975
       P(X >  1.95996398454005) = 0.025
```




```
[1] -1.959964  1.959964
```

CI Summary: Single Mean

To compute a confidence interval for a mean when the sampling distribution for $\bar{x}$ is approximately Normal (i.e., Normal population, or "large" n) and σ is unknown (which is almost always), use

$$\bar{x} \pm t^*_{n-1} \cdot \frac{s}{\sqrt{n}}$$

where t^*_{n-1} is the quantile appropriate for the confidence level, computed from a t-distribution with $n-1$ degrees of freedom.

Atlanta Commute Time: Analytic CI

- Confidence interval

$$\bar{x} \pm T^* \cdot \hat{SE}$$

- $\bar{x} = 29.11$
- Get T^* using confidence level and $df = n - 2$

```
xqt(c(0.025, 0.975), df = 500 - 2)

     [1] -1.964739  1.964739
```

- $\hat{SE}$: $\frac{s}{\sqrt{n}}$

```
sd(~Time, data = CommuteAtlanta)  # Need to find s first

     [1] 20.71831
```

Outline

P-values for a sample mean

Computing P-values when the null sampling distribution is approximately Normal (i.e., Population is normal OR sample size is "large") and σ is unknown (which is almost always) is the reverse process:

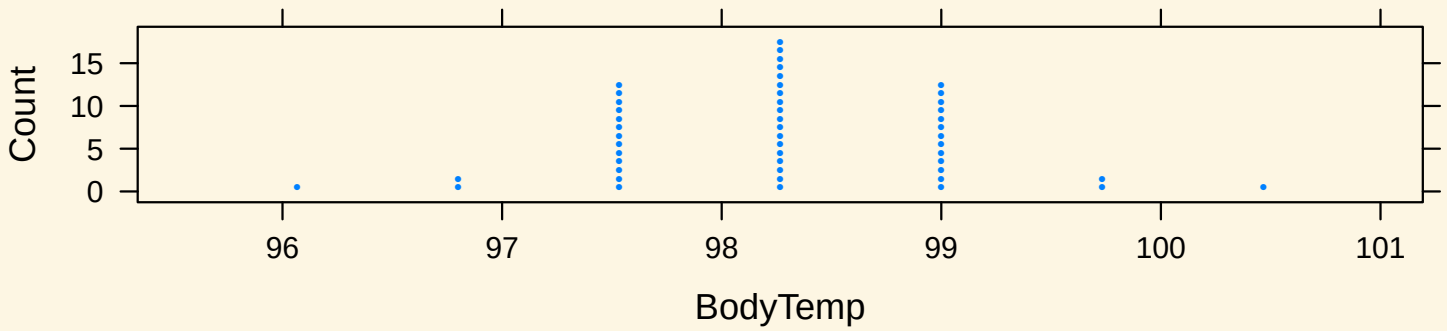
1. Convert $\bar{x}$ to a t-statistic within the theoretical distribution .

$$T_{observed} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

2. Find the relevant area beyond $T_{observed}$ using a t distribution with $n - 1$ degrees of freedom

Example: Mean Body Temperature

```
data("BodyTemp50")
dotPlot(~BodyTemp, data = BodyTemp50)
```




```
mean(~BodyTemp, data = BodyTemp50) # find the sample mean (x-bar)

     [1] 98.26

sd(~BodyTemp, data = BodyTemp50)   # find the sample sd (s)

     [1] 0.7653197
```

Example: Mean Body Temperature

- $H_0 : \mu = 98.6$
- Sample mean (standardized): $T_{obs} = \frac{\bar{x} - \mu_0}{\hat{SE}}$
- $\bar{x} = 98.26, \mu_0 = 98.6$
- $\widehat{SE} = \frac{s}{\sqrt{n}}$
- $s = 0.765, n = 50$
- Calculate t_{obs}

```
t.obs <- (98.26 - 98.6) / (0.765 / sqrt(50)); t.obs

     [1] -3.142697
```

- Once we have T_{obs}, find P-value from a t-distribution with $df = n - 1$

```
P.value <- 2 * xpt(-3.14, df = 50 - 1, lower.tail = TRUE); P.value

     [1] 0.002861716
```