Sampling Distributions

The most important distinction in statistics

sample
vs
population

When analysing data (or reading the literature), think about whether you want to discuss the sample that you observed or want to make statements that are more generally true

Statistics is the only field that gives us the correct framework to generalise from our sample to the population

The most important distinction in statistics

Example: T cell counts from 40 women with triple negative breast cancer were observed. What can we do with this information?

Option 1: Discuss the 40 women. What was the mean T cell count? What was its variation?

Option 2: Generalise the information about the 40 women to make statements about all women with triple negative breast cancer

These are 2 different approaches to using the same information

Language for making these distinctions

Population

- Size N (usually ∞)
- Mean = μ

$$\mu = \sum p_j X_j \quad or \quad \int \ldots$$

- Variance = σ^2

$$\sigma^2 = \sum p_j \left(X_j - \mu\right)^2 \quad or \quad \int \ldots$$

Sample

- Size n
- Sample Mean = $\overline{X}$

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

- Sample variance = s^2

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2$$

Generalising the sample to the population

Issue: We can calculate the sample mean and sample variance from our data, but the true mean and true variance are generally unknown

Fortunately, statisticians have learnt some things about how to recover* the true mean and true variance based only on sample means and sample variances

* with high probability

How do sample means behave?

$\bar{X}$

Suppose we observe data $X_1, X_2, ..., X_n$. We can calculate ~~Xbar~~ exactly, but what can we say about μ?

Idea: μ is probably close to Xbar

Goal: Make this more rigourous

Sums of Normal Random Variables

In general, neither $\Sigma_i X_i$ nor Xbar will have the same distribution as the X's

Example:
X_1, X_2, X_3 follow F-distributions.
$X_1+ X_2+ X_3$ does not follow an F-distribution
$(X_1+ X_2+ X_3)/3$ does not follow an F-distribution

Exception to the rule:
If $X_1, X_2, ..., X_n$ are independent and normally distributed with means μ_i and σ_i^2,

$$X_1+\ldots+X_n$$

follows a normal distribution with mean

$$\mu_1+\ldots+ \mu_n$$

and variance

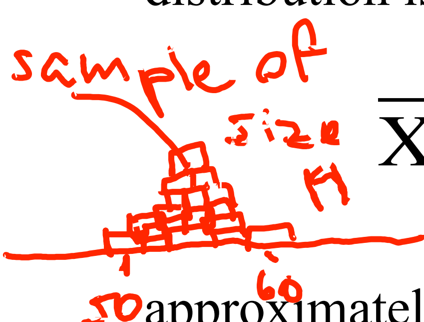
$$\sigma_1^2 +\ldots+ \sigma_n^2$$

Central limit theorem

What can we do instead? Use the central limit theorem!

variance

Central limit theorem:
If $X_1, X_2, \ldots, X_n$ are independent and have the same distribution, and the ~~mean~~ of that distribution is σ^2, then if n is large,



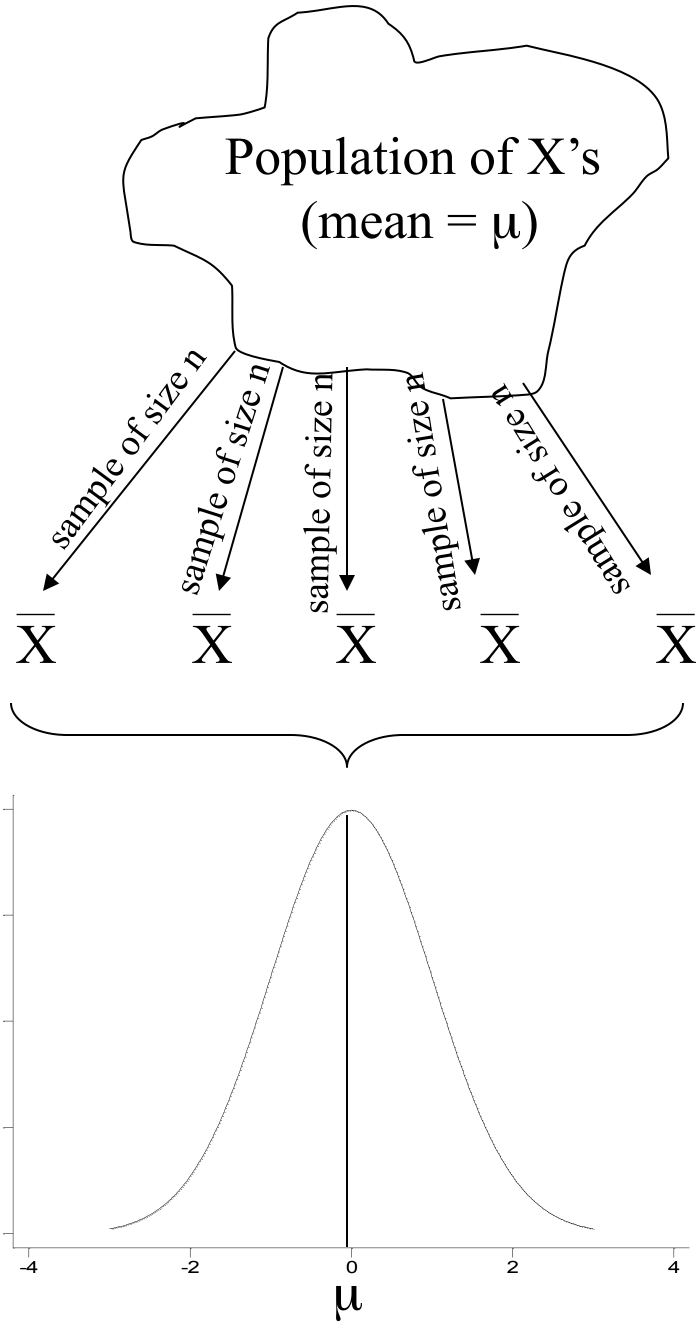
$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

approximately and under relatively weak conditions.

•In general, this applies for $n \geq 30$.

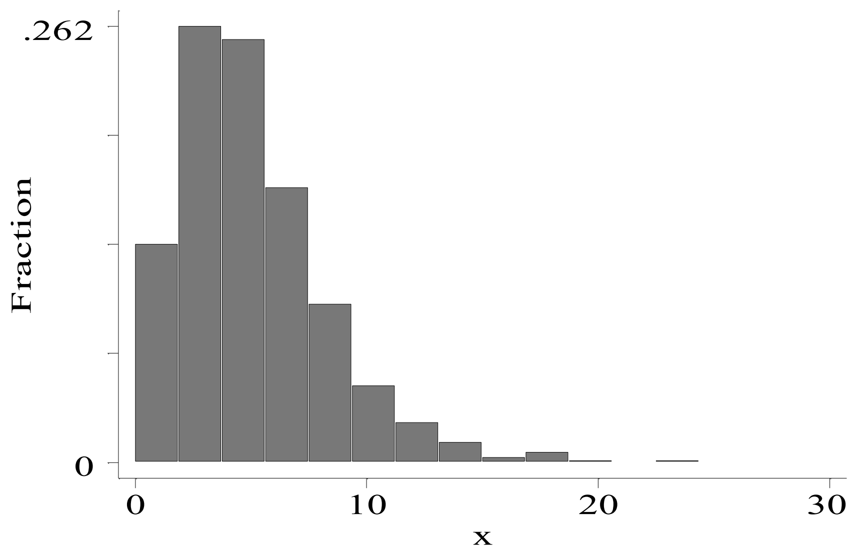
•As n increases, the normal approximation improves.

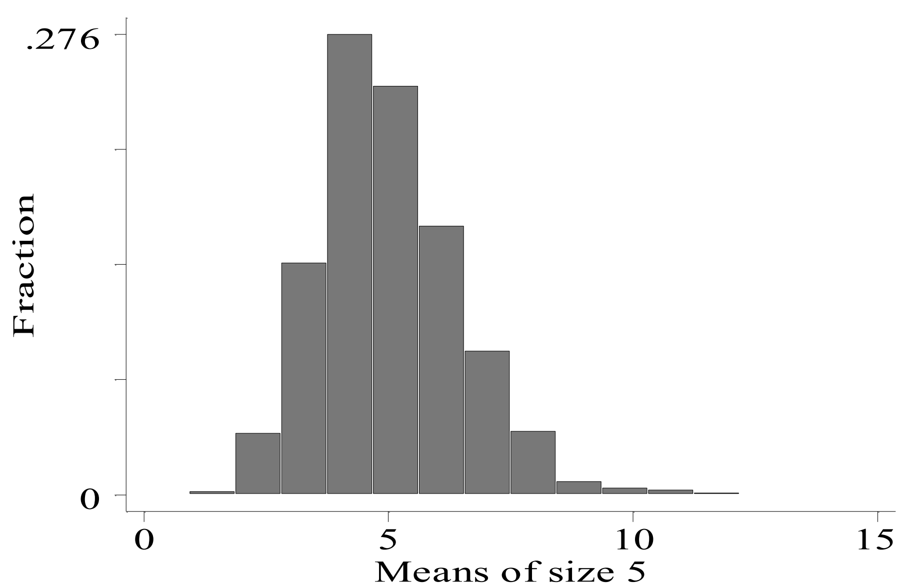
Distribution of the Sample Mean

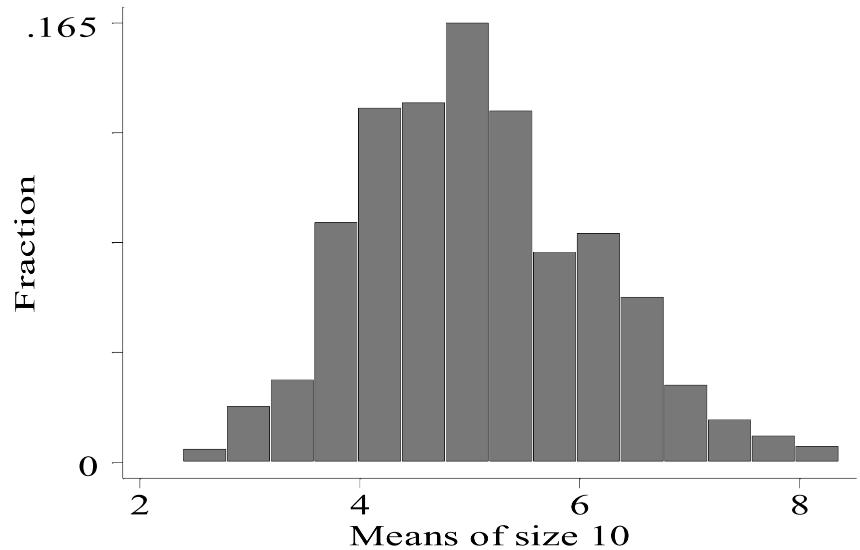


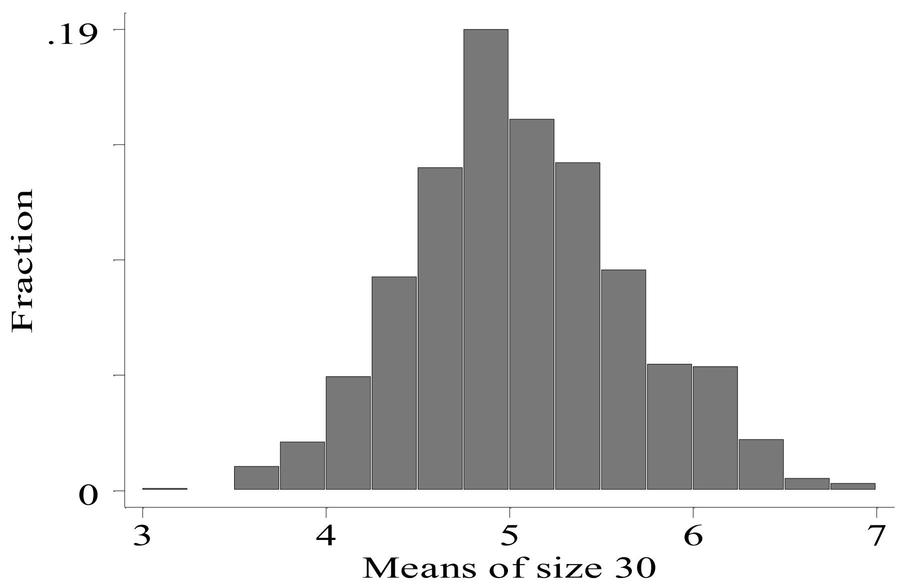
Central Limit Theorem - Illustration

Population






.165
Fraction
0
2
4
6
8
Means of size 10


.19
Fraction
0
3
4
5
6
7
Means of size 30

Central limit theorem

The central limit theorem allows us to use the sample ($X_1 \ldots X_n$) to discuss the population (μ)

We do not need to know the distribution of the data to make statements about the true mean of the population!

$X_1 \ldots X_n$

from some dist'n (?)

that has

mean μ (?)

var σ^2

known

Distribution of the Sample Mean

EXAMPLE:

Suppose that for Seattle sixth grade students the mean number of missed school days is 5.4 days with a standard deviation of 2.8 days. What is the probability that a random sample of size 49 (say Ridgecrest's 6th graders) will have a mean number of missed days greater than 6 days?

Find the probability that a random sample of size 49 from this population will have a mean greater than 6 days.

ill Seattle

$\mu = 5.4$ days

$\sigma = 2.8$ days → s.d. of # missed days for 1 child

$n = 49$

R.C.

$\sigma_{\overline{X}} = \sigma / \sqrt{n} = 2.8 / \sqrt{49} = 0.4$

$\mu_{\overline{X}} = 5.4$

$$P(\overline{X} > 6) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{6 - 5.4}{0.4}\right)$$

$$= P(Z > 1.5) = 0.0668$$

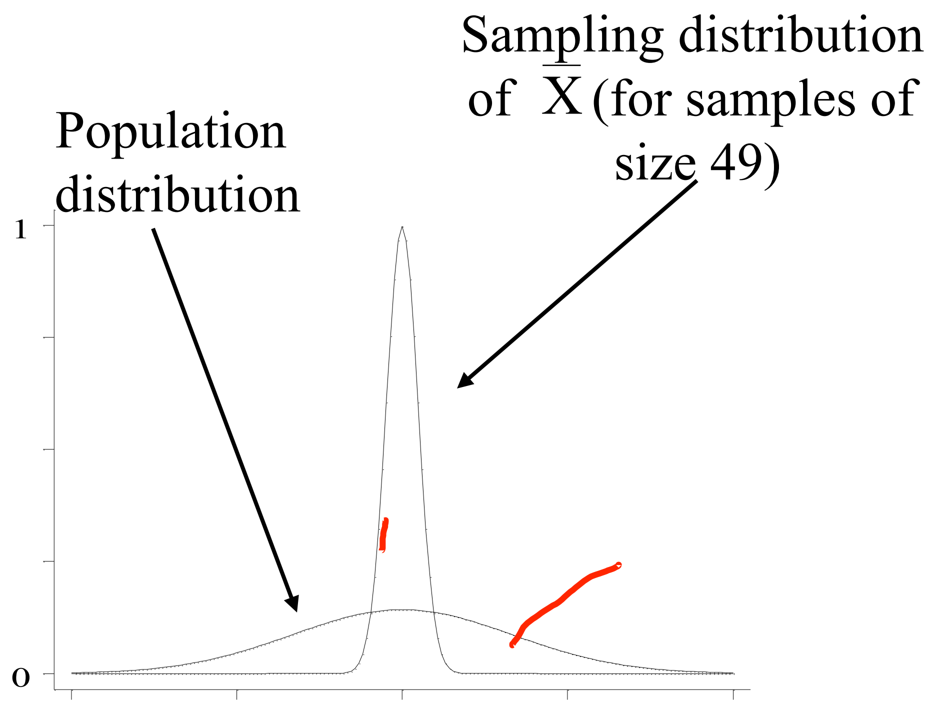
turn $N(5.4, \frac{\sigma^2}{n})$ into $N(0,1)$

$$Pr(\overline{X} > 6) = Pr\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{6 - 5.4}{2.8/\sqrt{49}}\right)$$

$$= Pr(Z > 1.5)$$

$N(0,1)$

$= .0668$



$$\mu = \mu_{\bar{X}} = 5.4$$

$X_i \sim ?$ mean μ

s.d σ

$\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$

s.s. n

Exercise

What is the probability that a random sample (size 49) from this population has a mean between 4 and 6 days?

Solution

$$
\begin{aligned}
P(4 \le \overline{X} \le 6) &= P(-3.5 \le Z \le 1.5) \\
&= P(Z \le 1.5) - P(Z \le -3.5) \\
&= .933
\end{aligned}
$$

Confidence Intervals

Confidence Intervals

Confidence intervals are not just intervals!

"(L, U) is a 100p% confidence interval for a parameter θ"

means that

"For any possible correct value parameter θ, the interval (L, U) contains θ with probability at least p."

Confidence intervals only concern parameters.

Prediction intervals (different!) are intervals about random variables.

Confidence Intervals for the mean

Because

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

we know that

$$P\left[-1.96 \le \underbrace{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}_{N(0,1)} \le +1.96\right] = 0.95.$$

Rearranging gives us that

$$\mu \in \left(\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval for the true mean μ

Confidence Intervals
σ known

If we desire a (1 - α) confidence interval we can derive it based on the statement

$$P\left[Q_Z^{\left(\frac{\alpha}{2}\right)} < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < Q_Z^{\left(1-\frac{\alpha}{2}\right)}\right] = 1 - \alpha$$

That is, we find constants $Q_Z^{\left(\frac{\alpha}{2}\right)}$ and $Q_Z^{\left(1-\frac{\alpha}{2}\right)}$ that have exactly (1 - α) probability between them.

A (1 - α) Confidence Interval for the Population Mean

$$\left(\overline{X} + Q_Z^{\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + Q_Z^{\left(1-\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Intervals
σ known - EXAMPLE

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of 30 second time mothers yield a mean pregnancy length of 279.5 days. Construct a 90% confidence interval for the mean length of second pregnancies based on this sample.

$$279.5 \pm Q_Z^{0.975} \times \frac{6}{\sqrt{30}}$$

$$279.5 \pm 1.645 \times \frac{6}{\sqrt{30}}$$

Confidence Intervals
σ unknown ← overwhelming

To get a CI for μ using the methods outlined above, we need $\bar{X}$ and σ^2. But usually, **σ is unknown** - we only have $\bar{X}$ and s^2. It turns out that even though

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

is normally distributed,

$$\frac{(\bar{X}-\mu)}{s/\sqrt{n}}$$

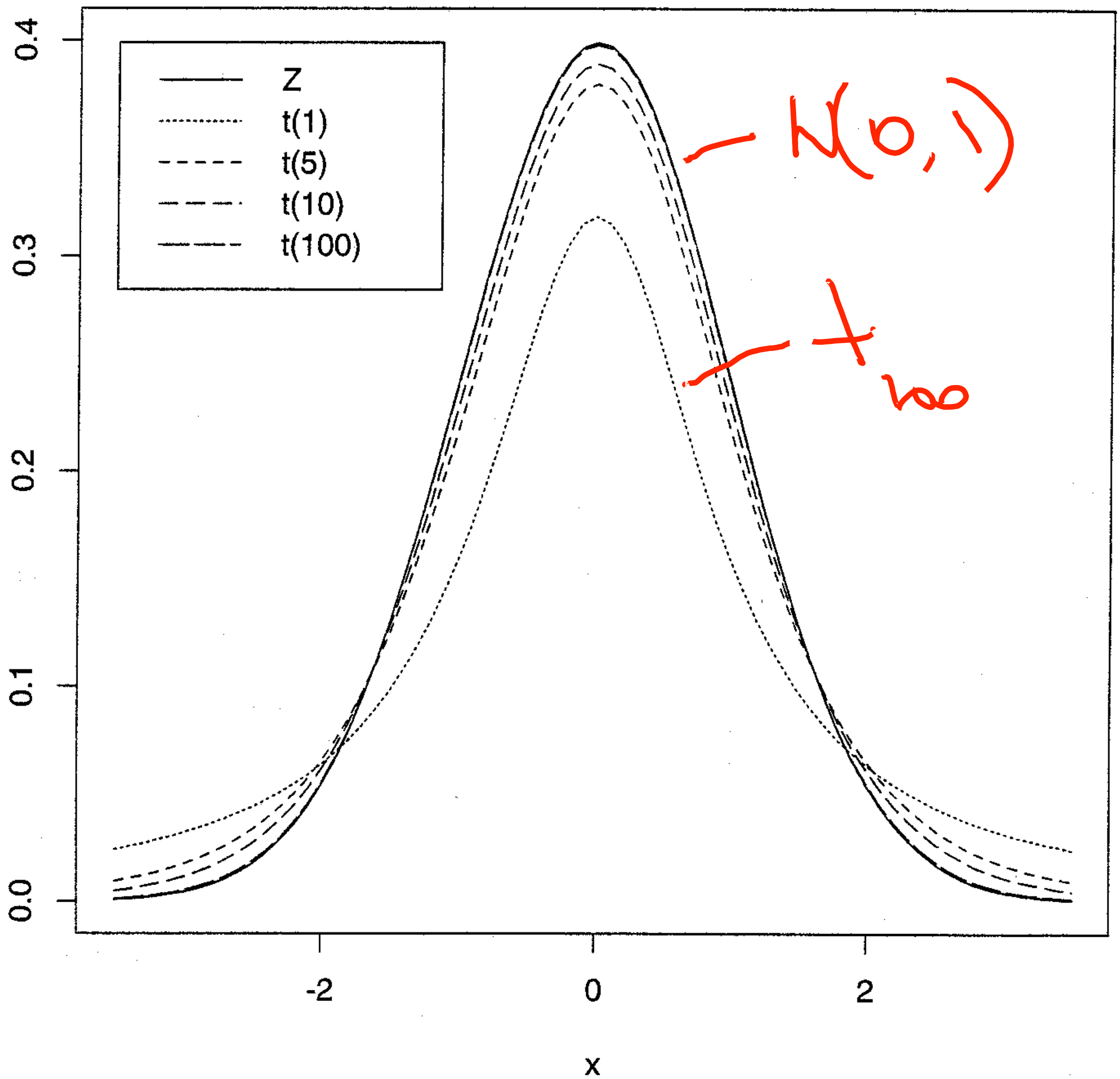
is not (quite)!

W.S. Gosset worked for Guinness Brewing in Dublin, IR. He was forced to publish under the pseudonym "Student". In 1908 he derived the distribution of

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

which is now known as Student's **t-distribution**.

Normal and t distributions



Confidence Intervals
σ² unknown
t Distribution

When σ is unknown we replace it with the estimate, s, and use the t-distribution. The statistic

$$\frac{\overline{X}-\mu}{s/\sqrt{n}}$$

has a t-distribution with n-1 *degrees of freedom*.

We can use this distribution to obtain a confidence interval for μ even when σ is not known.

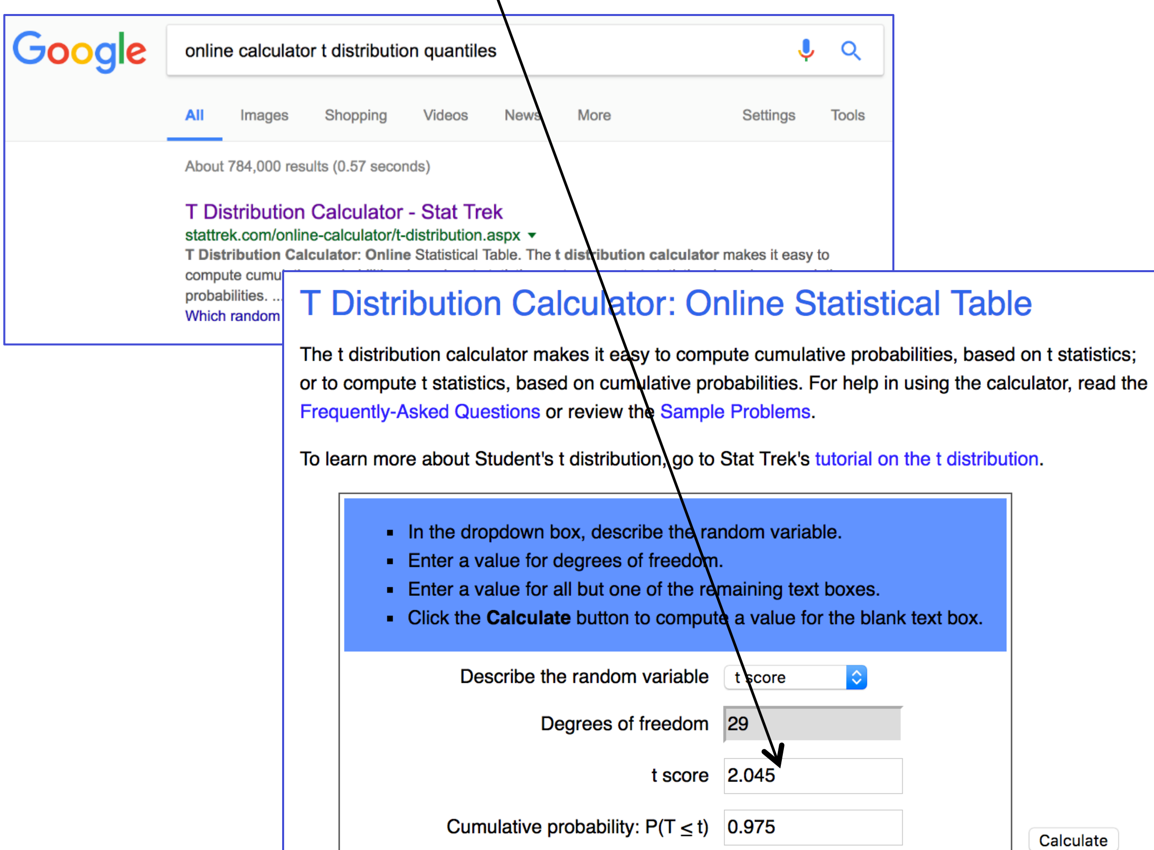
A (1-α) **Confidence Interval for the Population Mean when σ is unknown**

$$\left(\overline{X}+t_{(n-1)}^{\left(\frac{\alpha}{2}\right)}\times s/\sqrt{n},\ \overline{\mathrm{X}}+t_{(n-1)}^{\left(1-\frac{\alpha}{2}\right)}\times s/\sqrt{n}\right)$$

Confidence Intervals - σ^2 unknown
t Distribution - EXAMPLE

Given our 30 moms with a mean gestation of 279.5 days and a variance of 28.3 days^2, we can now compute a 95% confidence interval for the mean length of pregnancies for second time mothers:

$$279.5 \pm t_{29}^{0.975} \times \frac{\sqrt{28.3}}{\sqrt{30}}$$



Confidence Intervals - sample variance

Q: Can we derive a confidence interval for the sample variance?

A: Yes. We'll need the **Chi-square distribution**

Definition: The sum of squared independent standard normal random is a random variable with a **Chi-square** distribution with n degrees of freedom.

Let Z_i be standard normals, N(0,1). Let

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

X has a χ^2(n) **distribution**

Chi-square Distribution

Properties of χ^2 (n): Let $X \sim \chi^2(n)$.

1. $X \geq 0$
2. $E[X] = n$
3. $V[X] = 2n$
4. **n**, the parameter of the distribution is called *the degrees of freedom*.

Chi-square Distribution
Sample Variance

The Chi-square distribution describes the distribution of the **sample variance**. Recall

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

and

$$(n-1)\frac{s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

Now the right side almost looks like

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

which would be $\chi^2(n)$.

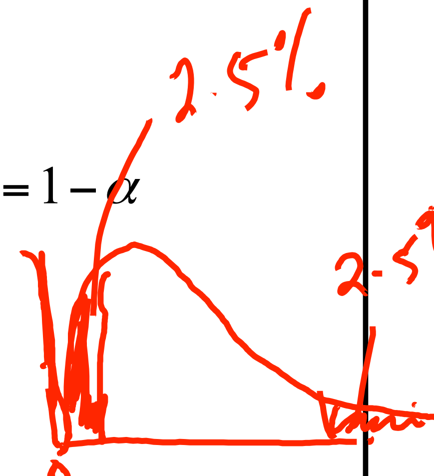
Since μ is estimated by $\overline{X}$ one degree of freedom is lost leading to …

$$\boxed{(n-1)\frac{s^2}{\sigma^2} \sim \chi^2}$$ with n-1 degrees of freedom

Chi-square Distribution
Confidence Interval for σ^2

We can use the Chi-square distribution to obtain a (1 - α) confidence interval for the **population variance**.

$$P\left[Q^{\left(\frac{\alpha}{2}\right)}_{\chi^2(n-1)} < (n-1)\frac{s^2}{\sigma^2} < Q^{\left(1-\frac{\alpha}{2}\right)}_{\chi^2(n-1)}\right] = 1-\alpha$$



Now, inverting this statement yields:

$$P\left[s^2 \times (n-1)/Q^{\left(1-\frac{\alpha}{2}\right)}_{\chi^2(n-1)} < \sigma^2 < s^2 \times (n-1)/Q^{\left(\frac{\alpha}{2}\right)}_{\chi^2(n-1)}\right] = 1-\alpha$$

Therefore,

A (1 - α) Confidence Interval for the Population Variance

$$\boxed{\left(s^2 \times (n-1)/Q^{\left(1-\frac{\alpha}{2}\right)}_{\chi^2(n-1)},\ s^2 \times (n-1)/Q^{\frac{\alpha}{2}}_{\chi^2(n-1)}\right)}$$

Chi-square Distribution
Confidence Interval for σ^2 - Exercise

Suppose for the second time mothers were not happy using the standard deviation of 6 days since it was based on the population of all mothers regardless of parity. The sample variance was 28.3 days^2. What is a 95% confidence interval for the variance of the length of second pregnancies?

Chi-square Distribution
Confidence Interval for σ^2 - Solution

Suppose for the second time mothers were not happy using the standard deviation of 6 days since it was based on the population of all mothers regardless of parity. The sample variance was 28.3 days^2. What is a 95% confidence interval for the variance of the length of second pregnancies?

$$n = 30, s^2 = 28.3$$
$$Q^{0.025}_{\chi^2_{29}} = 16.1$$
$$Q^{0.975}_{\chi^2_{29}} = 45.7$$
$$(28.3 \times \frac{29}{45.7}, 28.3 \times \frac{29}{16.1})$$
$$\Longrightarrow\ (17.96, 50.98) \text{ is a } 95\% \text{ CI for } \sigma^2$$

$$(4.24, 7.14) \text{ is a } 95\% \text{ CI for } \sigma$$

Summary

- General $(1 - \alpha)$ Confidence Intervals.
  - Confidence intervals are only for parameters!
- CI for μ, σ assumed known → Z.
- CI for μ, σ unknown → T.
- CI for $\sigma^2 \rightarrow \chi^2$

- ↑confidence → wider interval
- ↑sample size → narrower interval