Lecture 11 - Power, Inference for Means with Small Samples

Sta102 / BME102

Colin Rundel

October 6th, 2014

Example - Blood Pressure

Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg.

We are interested in finding out if the average blood pressure of employees at a certain company is greater than the national average, so we collect a random sample of 100 employees and measure their systolic blood pressure. What are the hypotheses?

We'll start with a very specific question – "What is the power of this hypothesis test to correctly detect an increase of 2 mmHg in average blood pressure?"

Calculating power

The preceeding question can be rephrased as – How likely is it that this test will reject H_0 when the true average systolic blood pressure for employees at this company is 132 mmHg?

Let's break this down intro two simpler problems:

1. Problem 1: Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
2. Problem 2: What is the probability that we would reject H_0 if $\bar{x}$ had come from $N\left(\mu = 132, \sigma = \frac{25}{\sqrt{100}} = 2.5\right)$, i.e. what is the probability that we can obtain such an $\bar{x}$ from this distribution?

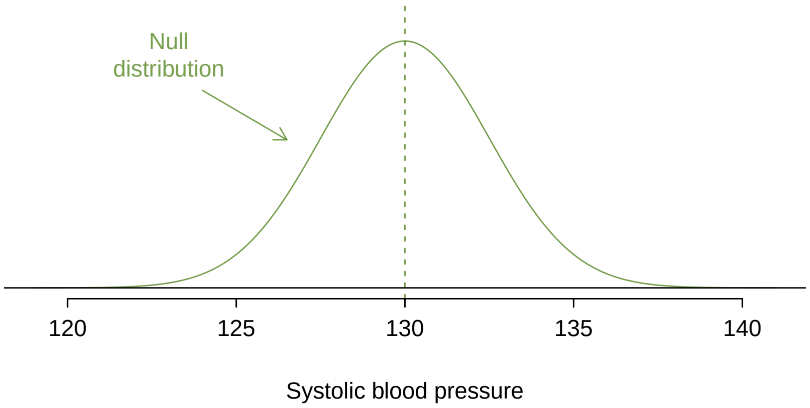
Problem 1

Which values of $\bar{x}$ represent sufficient evidence to reject H_0?
(Remember $H_0 : \mu = 130$, $H_A : \mu > 130$)

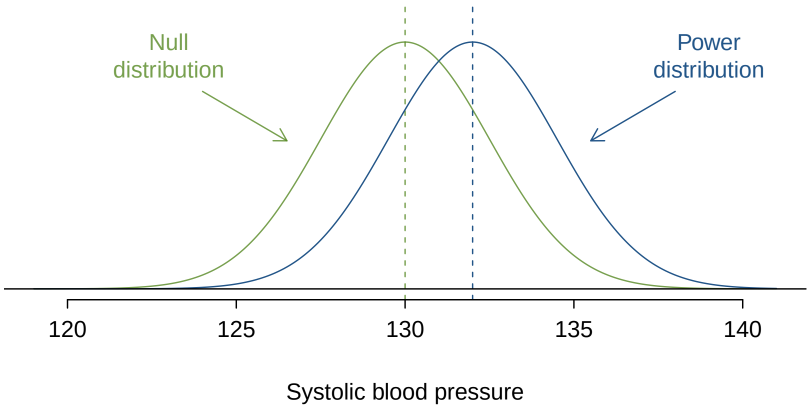
Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ did come from $N(\mu = 132, \sigma = 2.5)$.

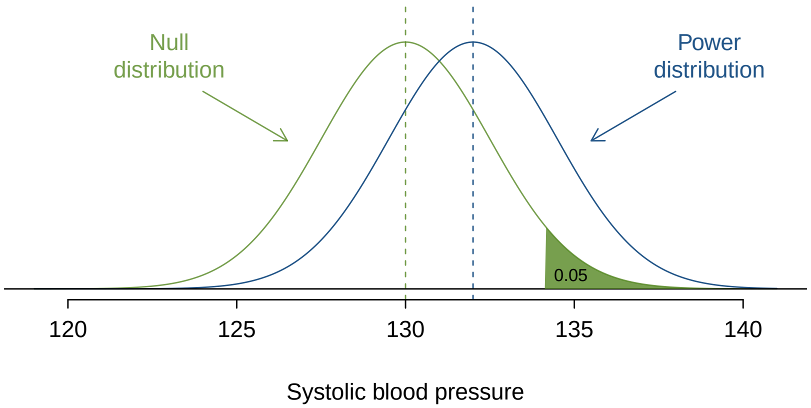
Putting it all together



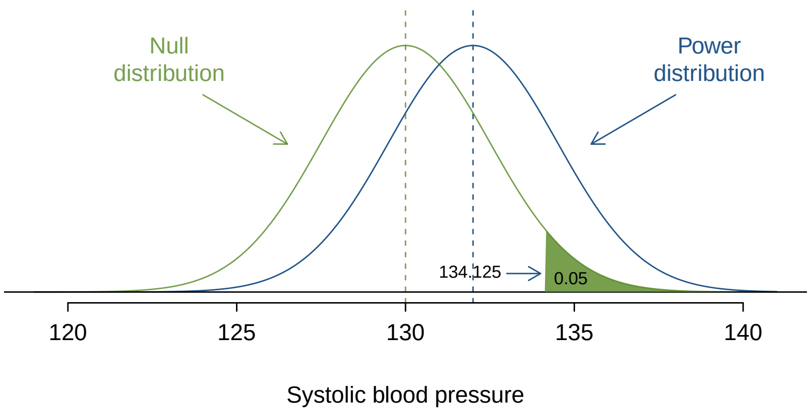
Putting it all together



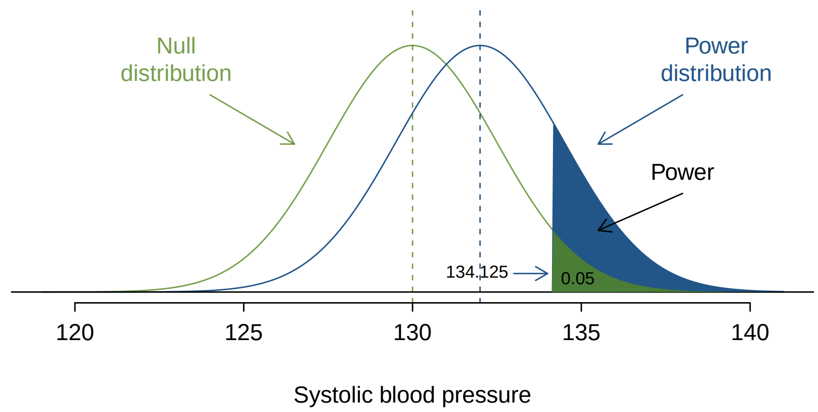
Putting it all together



Putting it all together



Putting it all together



Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.

Recap - Calculating Power

- Begin by picking a meaningful effect size δ and a significance level α
- Calculate the range of values for the point estimate beyond which you would reject H_0 at the chosen α level.
- Calculate the probability of observing a value from preceding step if the sample was derived from a population where $\bar{x} \sim N(\mu_{H_0} + \delta, SE)$

Example - Calculating power for a two sided hypothesis test

Going back to the blood pressure example, what would the power be to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level for a sample of 625 patients?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad n = 625, \quad \sigma = 25, \quad \delta = 4, \quad 1-\beta = ?$$

Step 1:

$$P(Z > z \text{ or } Z < -z) < 0.05 \quad \Rightarrow \quad z > 1.96$$
$$\bar{x} > 130 + 1.96\frac{25}{\sqrt{625}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{625}}$$
$$\bar{x} > 131.96 \text{ or } \bar{x} < 128.04$$

Step 2:

$$\bar{x} \sim N(\mu+\delta, SE) = N(134, 1)$$
$$P(\bar{x} > 131.96 \text{ or } \bar{x} < 128.04) = P(Z > [131.96 - 134]/1) + P(Z < [128.04 - 134]/1)$$
$$= P(Z > -2.04) + P(Z < -5.96) = 0.979 + 0 = 0.979$$

Example - Using power to determine sample size

Going back to the blood pressure example, how large a sample would you need if you wanted 90% power to detect a 4 mmHg increase in average blood pressure for the hypothesis that the population average is different from 130 mmHg at a 95% significance level?

Step 0:

$$H_0 : \mu = 130, \quad H_A : \mu \neq 130, \quad \alpha = 0.05, \quad \beta = 0.10, \quad \sigma = 25, \quad \delta = 4, \quad n = ?$$

Step 1:

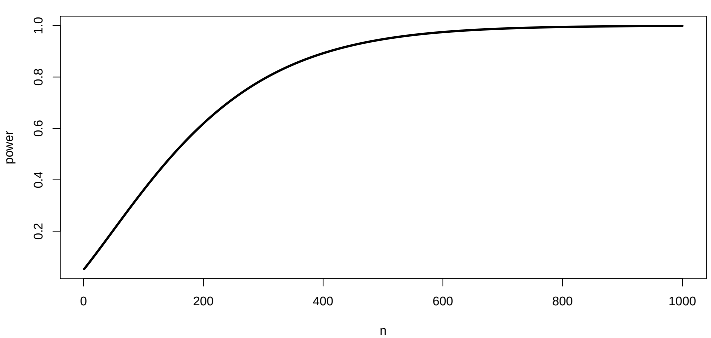
$$P(Z > z \text{ or } Z < -z) < 0.05 \quad \Rightarrow \quad z > 1.96$$
$$\bar{x} > 130 + 1.96\frac{25}{\sqrt{n}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{n}}$$

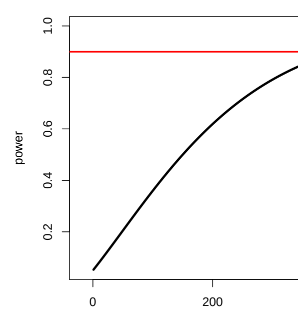
Step 2:

$$\bar{x} \sim N(\mu+\delta, SE) = N(134, 25/\sqrt{n})$$
$$P\left(\bar{x} > 130 + 1.96\frac{25}{\sqrt{n}} \text{ or } \bar{x} < 130 - 1.96\frac{25}{\sqrt{n}}\right) = 0.9$$
$$P\left(Z > 1.96 - 4\frac{\sqrt{n}}{25} \text{ or } Z < -1.96 - 4\frac{\sqrt{n}}{25}\right) = 0.9$$

Example - Using power to determine sample size (cont.)

So we are left with an equation we cannot solve directly, how do we evaluate it?

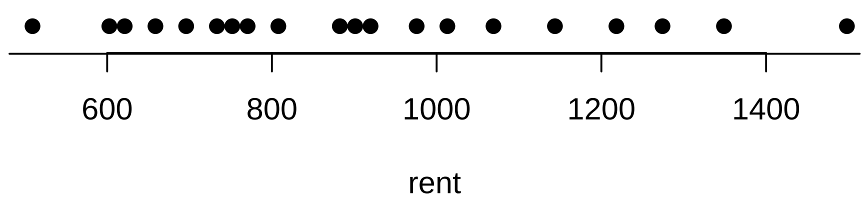




For $n = 410$ the power $= 0.8996$, therefore we need 411 subjects in our sample to achieve the desired level of power for the given circumstance.

Example - Rent in Durham

20 apartments here in Durham were randomly sampled and their rents obtained. The dot plot below shows the distribution of the rents of these apartments. Can we apply the methods we have learned so far to construct a confidence interval using these data. Why or why not?



Bootstrapping

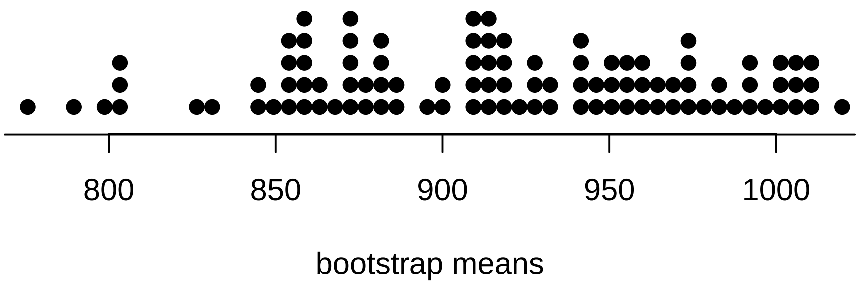
- An alternative approach to constructing confidence intervals is *bootstrapping*.
- This term comes from the phrase "pulling oneself up by one's bootstraps", which is a metaphor for accomplishing an impossible task without any outside help.
- In this case the ~~im~~possible task is estimating a population parameter, and we'll accomplish it using data from only the given sample.

Bootstrapping

- Bootstrapping works as follows:
  (1) take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  (2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  (3) repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- A bootstrap confidence interval is estimated by using the cutoff values for the middle 95% (or whatever other CL desired) of the bootstrap distribution.

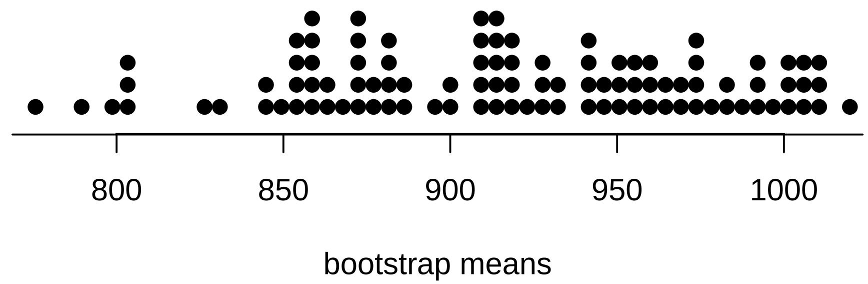
Example - Rent in Durham - Bootstrap interval

The dot plot below shows the distribution of means of 100 bootstrap samples from the original sample. We want to estimate the 95% bootstrap confidence interval based on this bootstrap distribution.



Example - Rent in Durham - Bootstrap interval

The dot plot below shows the distribution of means of 100 bootstrap samples from the original sample. We want to estimate the 90% bootstrap confidence interval based on this bootstrap distribution.

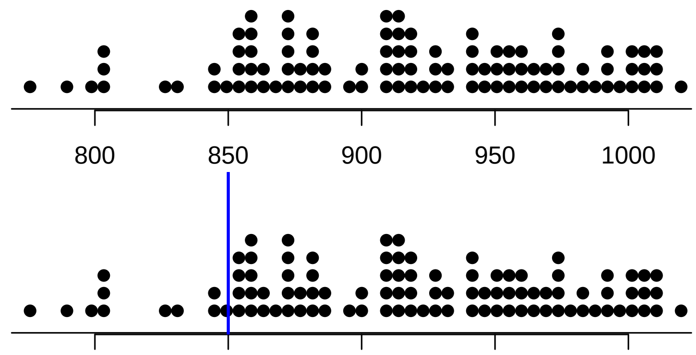


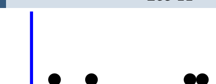
Randomization testing for a mean

- We can also use a simulation method to conduct the equivalent of a hypothesis test.
- This is very similar to bootstrapping, i.e. we once again randomly sample with replacement from the sample, but this time we calculate a p-value using the bootstrap distribution.
- As with a theoretical hypothesis test, the p-value is defined as the proportion of simulations that yield a sample mean at least as favorable to the alternative hypothesis as the observed sample mean.

Example - Rent in Durham - Randomization testing

You read an article claiming that the average rent for an apartment in Durham is \$850. The original random sample had a mean of \$920.1. Does this sample provide convincing evidence that the article's estimate is an underestimate?



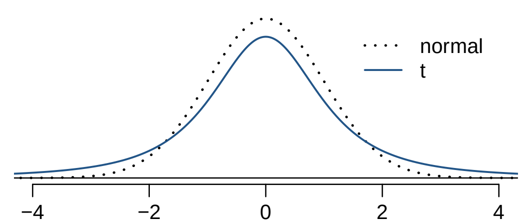


What purpose does a large sample serve?

- As long as observations are independent a large sample ensures that the sample average will have a nearly normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
- But when it comes to inference there is a problem, we almost never know σ.
- If n is large enough then s should be close to σ just like $\bar{X}$ is close to μ.
- What do we do when n isn't large then? *Use a more conservative distribution*

The *t* distribution

- When working with small samples, and the population standard deviation is unknown (this is almost always the case), the uncertainty of the standard error estimate is addressed by using a new distribution - the *t distribution*.
- This distribution also is also bell shape, but its tails are *thicker* than the normal.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate of the standard error (since n is small)



History of the t distribution

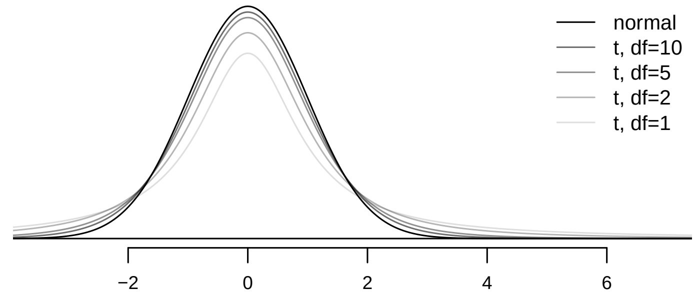
First published by by William Gosset ...

- Oxford Graduate with a degree in Chemistry and Mathematics
- Hired as a brewer by the Guinness Brewery in 1899
- Spent 1906 - 1907 studying with Karl Pearson
- Published "The probable error of a mean" in 1908 under the pseudonym "Student"
- Much of his work was promoted by R.A. Fisher

Properties of the t distribution

The t distribution ...

- is always centered at zero, like the standard normal (Z) distribution.
- has a single parameter: *degrees of freedom* (*df*).



- as *df* increases the t distribution converges to the unit normal distribution.

Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 30$) mean is the T statistic with $df = n - 1$.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

In context of the Durham rent data ...

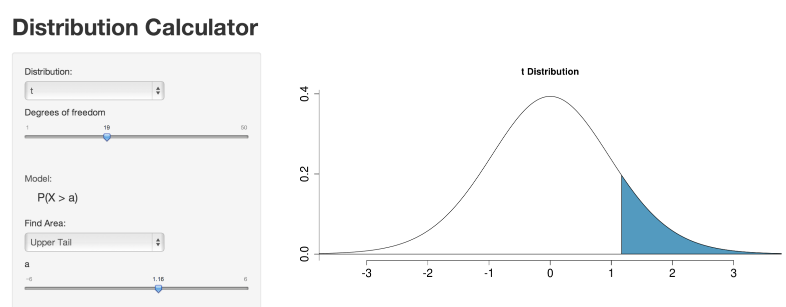
$$\begin{aligned}
\textit{point estimate} &= \overline{X} = 920.1 \\
SE &= \frac{s}{\sqrt{n}} = \frac{271}{\sqrt{20}} = 60.6 \\
T &= \frac{920.1 - 850}{60.6} = 1.16 \\
df &= 10 - 1 = 9
\end{aligned}$$

Finding the p-value

- As before the p-value is calculated as the area under the tail of a distribution, but here we use the t distribution.
- Using R:

```
1-pt(1.16,df=19)
## [1] 0.1302
```

- Using a web applet (`http://bit.ly/dist_calc`):



Finding the p-value - t table

Locate the T statistic on the appropriate df row, obtain the p-value from the corresponding column heading (one or two tail, depending on the alternative hypothesis).

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮	⋮	⋮	⋮	⋮	
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

Finding the p-value (cont.)

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85

$$T = 0.82 \quad \text{p-value} > 0.10$$

What is the conclusion of the hypothesis test?

Confidence interval for a small sample mean

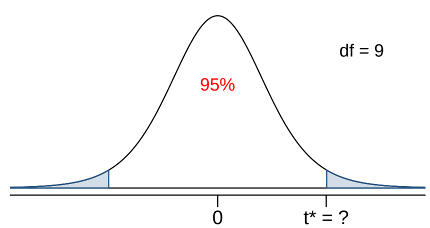
- Confidence intervals are always of the form

$$\text{point estimate} \pm ME$$

- ME is always calculated as the product of a critical value and SE.
- Since small sample means follow a t distribution (and not a Z distribution), the critical value is a $t^\star$ (as opposed to a $Z^\star$).

$$\text{point estimate} \pm t^\star_{df} \times SE$$

Finding the critical t ($t^\star$)



$n = 10, \; df = 10 - 1 = 9$

$t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the average rental price of an apartment in Durham?

$$\overline{X} = 920.1 \qquad s = 271 \qquad n = 10 \qquad SE = 60.6$$

Recap: Inference using a small sample mean

- If $n < 30$, and σ is unknown, then $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ has a t distribution with $df = n - 1$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - $n < 30$, population not overly skewed
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$