BME 1132
Probability and Biostatistics

Instructor: Ali AJDER, *Ph.D.*

Week-11

- Introduction
- The Relationship Between Population and Sample
- Estimation of the Mean of a Distribution
- Estimation of the Variance of a Distribution
- The Law of Large Numbers
- Central- Limit Theorem
- Confidence Intervals

Introduction-1

Until now, we always assumed the specific probability distributions were known.

Infectious Disease We assumed the number of neutrophils in a sample of 100 white blood cells was binomially distributed, with parameter $p = .6$.

Bacteriology We assumed the number of bacterial colonies on a 100-cm^2 agar plate was Poisson distributed, with parameter $\mu = 2$.

Hypertension We assumed the distribution of diastolic blood-pressure (DBP) measurements in 35- to 44-year-old men was normal, with mean $\mu = 80$ mm Hg and standard deviation $\sigma = 12$ mm Hg.

The only question that was what we can predict about the behavior of the data given an understanding of these properties.

Introduction-2

The problem is that we have a data set and we want to **infer** the properties of the underlying distribution from this data set. This inference usually involves **inductive reasoning** rather than **deductive reasoning**.

We are specifically interested in population mean and population variance of a random variable. These quantities are unknown in general. We refer to these unknown quantities as **parameters**.

We will discuss statistical methods for parameter **estimation** in this week.
Estimation refers to the process of guessing the unknown value of a parameter (e.g., population mean) using the observed data. For this, we will use an **estimator**, which is a **statistic**.

Sometimes we only provide a single value as our estimate. This is called **point estimation**. Point estimates do not reflect our uncertainty when estimating a parameter. To address this issue, we can present our estimates in terms of a range of possible values (as opposed to a single value). This is called **interval estimation**.

Estimation of the Population Mean

For a population of size N, μ is calculated as

$$\mu = \sum_{i=1}^{N} x_i / N$$

A natural estimator to use for estimating the population mean μ is the sample mean

$$\bar{X} = \sum_{i=1}^{n} X_i / n$$

In this case, we say that $\bar{X}$ is an estimator for μ .

What properties of $\bar{X}$ make it a desirable estimator of μ?

Note: We must forget about our particular sample for the moment and consider the set of all possible samples of size *n* that could have been selected from the population.

Example

The study by Mackowiak et al. aimed at estimating the population mean for body temperature among healthy people. From a sample of n = 148 people, they estimated the unknown population mean with the sample mean $\bar{\mu} = \bar{x} = 98.25°\text{F} = 36.8\,°\text{C}$. This estimate is lower than the commonly believed value of $98.6\,°\text{F} = 37\,°\text{C}$.

The sample size for this study was relatively small. We would expect that as the sample size increases, our point estimate based on the sample mean would become closer to the true population mean.

The Law of Large Numbers (LLN)

The **Law of Large Numbers** (LLN) indicates that (under some general conditions such as independence of observations) the sample mean converges to the population mean ($\bar{X}_n \rightarrow \mu$) as the sample size n increases ($n \rightarrow \infty$).

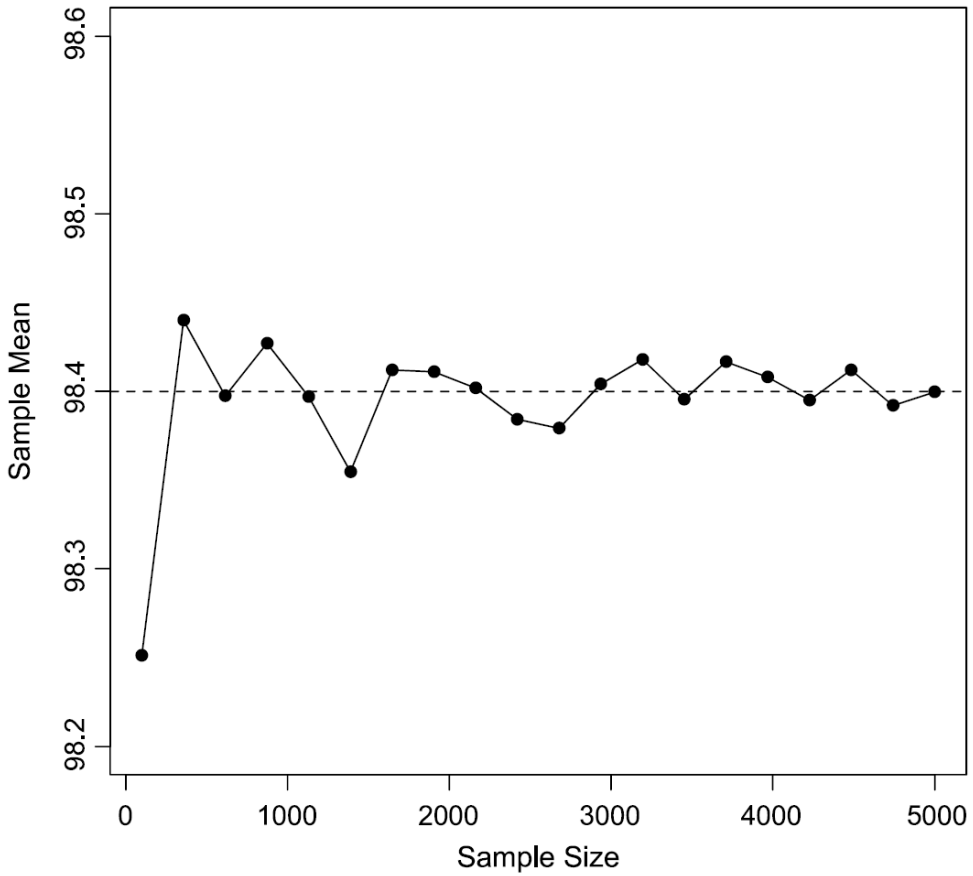
Informally, this means that the difference between the sample mean and the population mean tends to become smaller and smaller as we increase the sample size. The LLN provides a theoretical justification for the use of sample mean as an estimator for the population mean.

The LLN- Example

As an example, suppose that the true population mean for normal body temperature is 98.4°F.

As we gradually increase the sample size from 100 to 5000, the plot of the sample means (i.e., the point estimates of the population mean) might look like in Figure.

Here, the estimate of the population mean is plotted as a function of the sample size. As the sample size increases, the sample means converge to the true (but unknown) population mean $\mu = 98.4$.



Estimation of the Population Variance

The population variance, σ^2, is calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

A commonly used estimator for σ^2 is the sample variance,

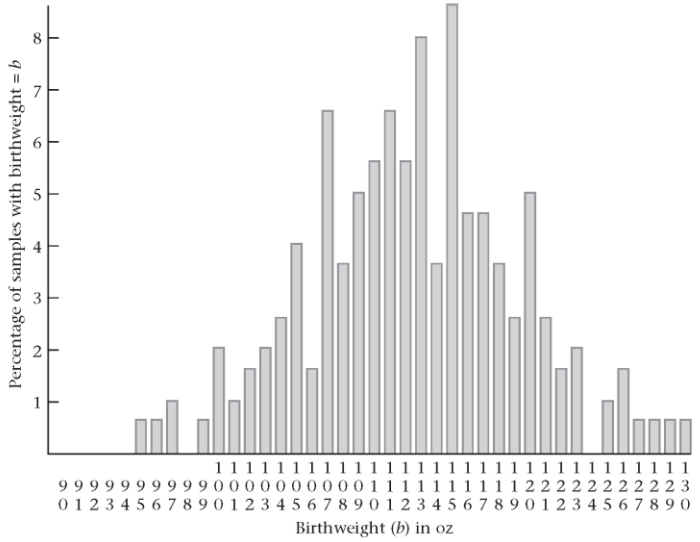
$$S^2 = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{n-1}$$

Note: Dividing by $n-1$ instead of n increases the value of the estimator by a small amount, which is enough to avoid underestimation associated with the more natural estimator.

Point Estimation

The **sampling distribution** of $\overline{X}$ is the distribution of values of $\overline{x}$ over all possible samples of size n that could have been selected from the reference population.

Figure gives an example of such a sampling distribution.
This is a frequency distribution of the sample mean from 200 randomly selected samples of size 10 drawn from the distribution of 1000 birthweights.



Note:
X : a random variable,
x : a specific realization of the random variable X in a sample.

Point Estimation

Let $X_1, \ldots, X_n$ be a random sample drawn from some population with mean μ.
Then, for the sample mean $\overline{X}$, $E(\overline{X}) = \mu$.

We refer to an estimator of a parameter θ as $\hat{\theta}$. An **estimator** $\hat{\theta}$ of a parameter θ is **unbiased** if $E(\hat{\theta}) = \theta$. This means that the average value of $\hat{\theta}$ over a large number of random samples of size n is θ.

The unbiasedness of $\bar{X}$ is **not sufficient reason** to use it as an estimator of μ.

The sample median and the average value of the largest and smallest data points in a sample are also unbiased.

The reason is that if the underlying distribution of the population is **normal**, then it can be shown that the unbiased estimator with the smallest variance is given by $\bar{X}$.
Thus, $\bar{X}$ is called the **minimum variance unbiased estimator** of μ.

Note that the above Equation holds for any population regardless of its underlying
distribution. In words, we refer to $\bar{X}$ as an unbiased estimator of μ.

Standard Error of the Mean

$\bar{X}$ is preferable to estimate parameters from large samples rather than from small ones, because of that the larger the sample size, the more precise an estimator $\bar{X}$ is.

Let $X_1, \ldots, X_n$ be a random sample from a population with underlying mean μ and variance σ^2. The set of sample means in repeated random samples of size n from this population has variance σ^2/n. The standard deviation of this set of sample means is thus $\sigma/\sqrt{n}$ and is referred to as the *standard error of the mean* or the *standard error.*

In practice, the population variance σ^2 is rarely known.
A reasonable estimator for the population variance σ^2 is the sample variance s^2, which leads to the following definition:

The **standard error of the mean (sem)**, or the **standard error (se)**, is given by $\sigma/\sqrt{n}$ and is estimated by $s/\sqrt{n}$. The standard error represents the estimated standard deviation obtained from a set of sample means from repeated samples of size n from a population with underlying variance σ^2.

Note:

$$Var(\bar{X}) = \left(\frac{1}{n^2}\right) Var\left(\sum_{i=1}^{n} X_i\right)$$

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

However, by definition $Var(X_i) = \sigma^2$. Therefore,

$$Var(\bar{X}) = (1/n^2)(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = (1/n^2)(n\sigma^2) = \sigma^2/n$$

The standard deviation $(sd) = \sqrt{\text{variance}}$; thus, $sd(\bar{X}) = \sigma / \sqrt{n}$.

Central- Limit Theorem

Central-Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. Then, for large n, $\overline{X} \dot{\sim} N\left(\mu, \sigma^2/n\right)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol $\dot{\sim}$ is used to represent "approximately distributed.")

This theorem is very important because many of the distributions encountered in practice are not normal. In such cases the central-limit theorem can often be applied; this lets us perform statistical inference based on the approximate normality of the sample mean despite the nonnormality of the distribution of individual observations.

Interval Estimation

We have assumed previously that the distribution of birthweights was normal with mean μ and variance σ^2. It follows from our previous discussion of the properties of the sample mean that $\bar{X} \sim N(\mu, \sigma^2/n)$. Thus, if μ *and* σ^2 were known then the behavior of the set of sample means over a large number of samples of size n would be precisely known.

In particular, 95% of all such sample means will fall within the interval $(\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n})$

Alternatively, if we re-express $\bar{X}$ in standardized form by

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

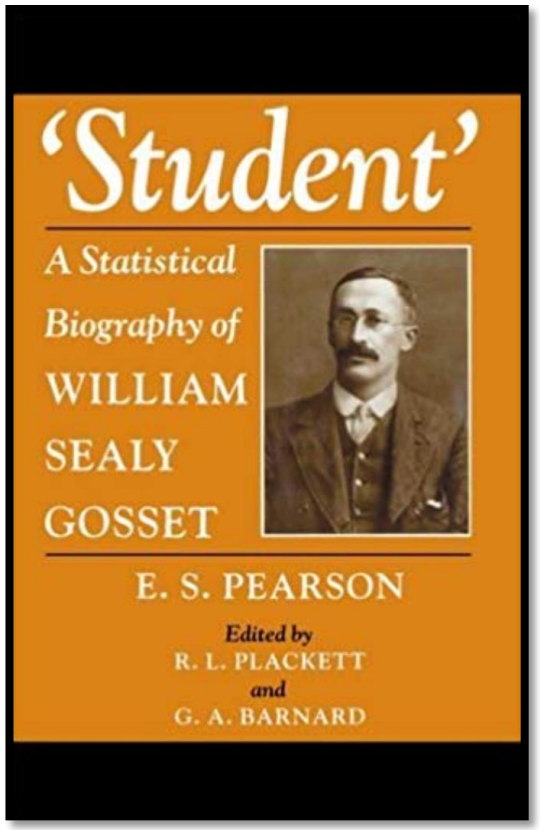
then Z should follow a standard normal distribution. Hence, 95% of the Z values from repeated samples of size n will fall between −1.96 and +1.96 because these values correspond to the 2.5th and 97.5th percentiles from a standard normal distribution. However, the assumption that σ is known is somewhat artificial, because σ is rarely known in practice.

t Distribution

Because **σ** is unknown, it is reasonable to estimate **σ** by the sample standard deviation ***s*** and to try to construct ***CI***s using the quantity $(\bar{\bar{X}} - \mu)/(S/\sqrt{n})$.

The problem is that this quantity is no longer normally distributed.

This problem was first solved in 1908 by a statistician named **William Gossett**. For his entire professional life, Gossett worked for the Guinness Brewery in Ireland. He chose to identify himself by the pseudonym "Student," and thus the distribution of $(\bar{\bar{X}} - \mu)/(S/\sqrt{n})$ is usually referred to as **Student's t distribution**. Gossett found that the shape of the distribution depends on the sample size n. Thus, the t distribution is not a unique distribution but is instead a family of distributions indexed by a parameter referred to as the **degrees of freedom** (df) of the distribution.

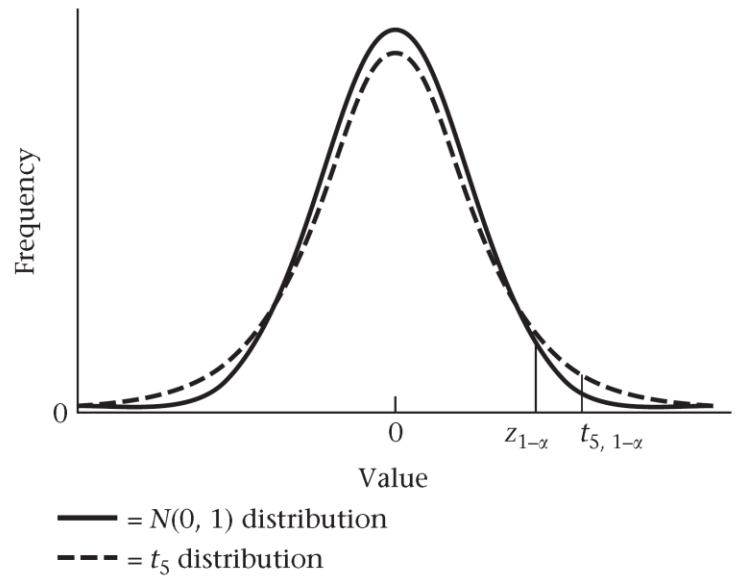


If $X_1, \ldots, X_n \sim N(\mu,\sigma^2)$ and are independent, then $(\bar{X} - \mu)/(S/\sqrt{n})$ is distributed as a t distribution with $(n - 1)$ *df*.

t Distribution

It is interesting to compare a ***t* distribution** with *d* degrees of freedom with an ***N*(0, 1)** distribution. The density functions corresponding to these distributions are depicted in Figure for the special case where $d = 5$.

Comparison of Student's *t* distribution with 5 degrees of freedom with an *N*(0, 1) distribution



The two distributions thus get more and more alike as *n* increases in size. The upper 2.5th percentile of the *t* distribution for various degrees of freedom and the corresponding percentile for the normal distribution are given in Table.

Comparison of the 97.5th percentile of the *t* distribution and the normal distribution

d	$t_{d,.975}$	$z_{.975}$	d	$t_{d,.975}$	$z_{.975}$
4	2.776	1.960	60	2.000	1.960
9	2.262	1.960	∞	1.960	1.960
29	2.045	1.960			

t Distribution

Confidence Interval for the Mean of a Normal Distribution

A 100% × (1 – α) CI for the mean μ of a normal distribution with unknown variance is given by

$$\left(\bar{x} - t_{n-1,1-\alpha/2}\, s/\sqrt{n}\,, \bar{x} + t_{n-1,1-\alpha/2}\, s/\sqrt{n}\right)$$

A shorthand notation for the CI is

$$\bar{x} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n}$$

Example

Compute a 95% CI for the following 10 birthweights.

2749.9 3543.7 1757.7 3401.9 3742.1 3827.2 3345.2 3883.9 3572 3345.2

Note:
Because we want a 95% CI, $\alpha = .05$

t Distribution

TABLE 5 Percentage points of the t distribution ($t_{d,u}$)[a]

Degrees of freedom, d	u								
	.75	.80	.85	.90	.95	.975	.99	.995	.9995
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

TABLE 5 Percentage points of the t distribution ($t_{d,u}$)[a]

Degrees of freedom, d	u								
	.75	.80	.85	.90	.95	.975	.99	.995	.9995
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

[a]The uth percentile of a t distribution with d degrees of freedom.

Source: Table 5 is taken from Table III of Fisher and Yates: "Statistical Tables for Biological, Agricultural and Medical Research," published by Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh).

t Distribution

Confidence Interval for the Mean of a Normal Distribution (Large-Sample Case)

An approximate $100\% \times (1 - \alpha)$ CI for the mean μ of a normal distribution with unknown variance is given by

$$\left(\bar{x} - z_{1-\alpha/2}\, s/\sqrt{n}, \bar{x} + z_{1-\alpha/2}\, s/\sqrt{n}\right)$$

Note:
This interval should only be used if $n > 200$.

In addition, this Equation can also be used for $n \leq 200$ if the standard deviation (σ) is **known**, by replacing s with σ.

Estimation of the Variance of a Distribution

The Chi-Square Distribution

To obtain an interval estimate for σ^2, a new family of distributions, called chi-square (χ^2) distributions, must be introduced to enable us to find the sampling distribution of S^2 from sample to sample.

If $G = \sum_{i=1}^{n} X_i^2$

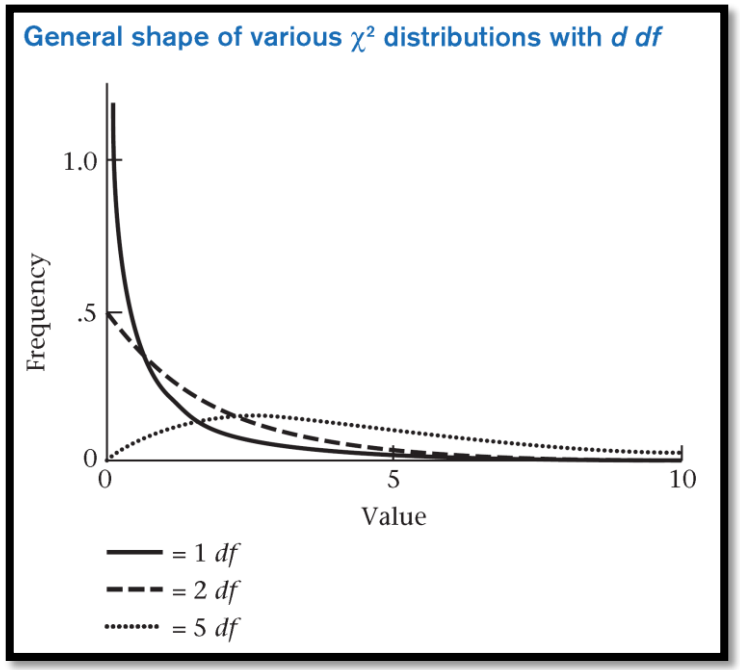
where $X_1, \ldots, X_n \sim N(0,1)$

and the $X_i's$ are independent, then G is said to follow a **chi-square distribution with n degrees of freedom (df)**. The distribution is often denoted by χ_n^2.

Estimation of the Variance of a Distribution

The Chi-Square Distribution

The chi-square distribution is actually a family of distributions indexed by the parameter ***n*** referred to, again, as the degrees of freedom, as was the case for the *t* distribution. Unlike the *t* distribution, which is always symmetric about 0 for any degrees of freedom, the chi-square distribution only takes on **positive values** and is always **skewed to the right**. The general shape of these distributions is indicated in Figure.



χ^2 Distribution

A $100\% \times (1 - \alpha)$ CI for σ^2 is given by

$$\left[(n-1)s^2/\chi^2_{n-1,1-\alpha/2}, (n-1)s^2/\chi^2_{n-1,\alpha/2}\right]$$

Example
Find the upper and lower **2.5th** percentiles of a chi-square distribution with **10** ***df***.

Example
Suppose we want to construct a 95% CI for the interobserver variability as defined by $\boldsymbol{\sigma^2}$.

χ^2 Distribution

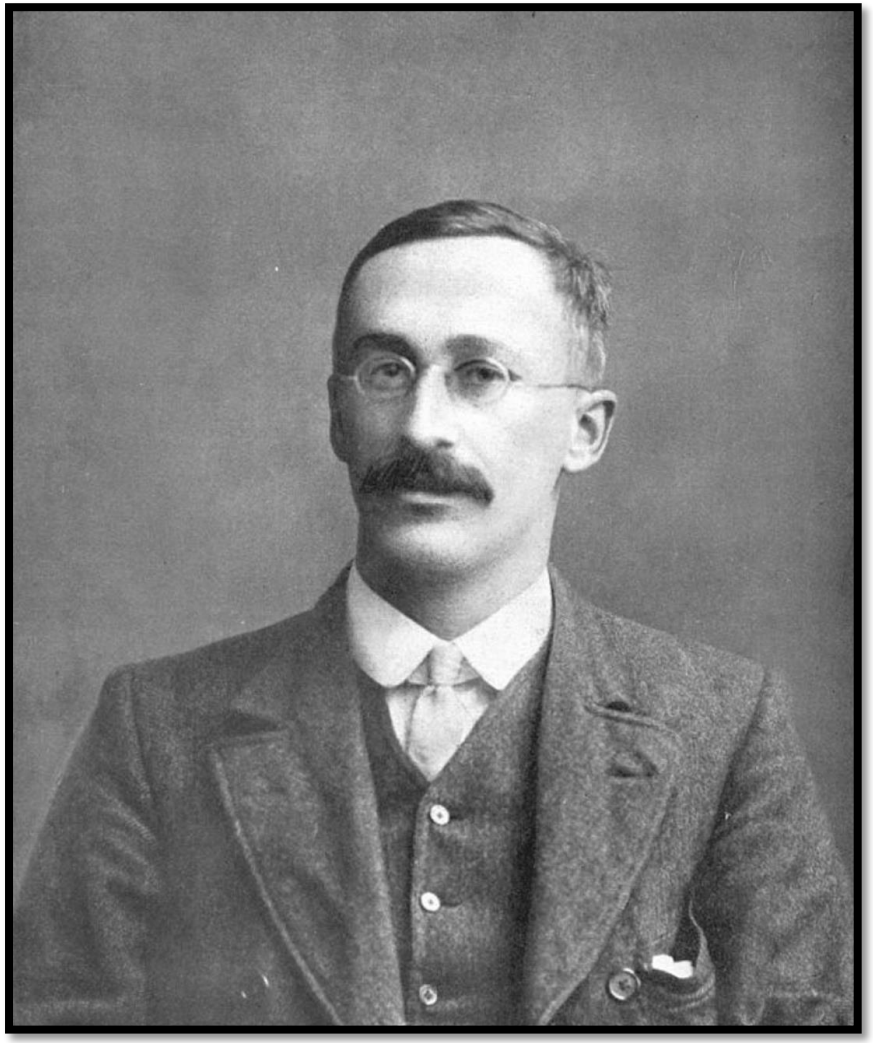
TABLE 6 Percentage points of the chi-square distribution ($\chi^2_{d,u}$)[a]

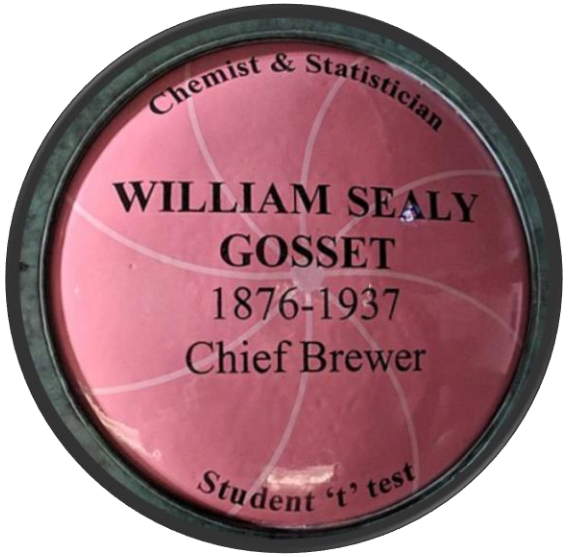
	u												
d	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995	.999
1	0.0^4393[b]	0.0^3157[c]	0.0^3982[d]	0.00393	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.0100	0.0201	0.0506	0.103	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.81
3	0.0717	0.115	0.216	0.352	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32

	u												
d	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995	.999
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

[a] $\chi^2_{d,u}$ = uth percentile of a χ^2 distribution with d degrees of freedom.
[b] = 0.0000393
[c] = 0.000157
[d] = 0.000982
Source: Based on the Biometrika Trustees, from Table 3 of *Biometrika Tables for Statisticians*, Volume 2, edited by E. S. Pearson and H. O. Hartley.

Questions?





William Sealy Gosset (13 June 1876 – 16 October 1937)
was Head Brewer of Guinness, Head Experimental Brewer of Guinness, and father of modern British statistics. He pioneered small sample experimental design and analysis with an economic approach to the logic of uncertainty.
Gosset published under the pen name "Student," and developed most famously ***Student's t-distribution*** *- originally called Student's "z" - and "Student's test of statistical significance".*[1]